STRICT ANSWER GRADING CAN REVERSE ITS ERROR DURING FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning models are trained, selected and compared by the accuracy of the last boxed answer they write, the default in widely used frameworks. We show that this rule can misreport progress by tens of points, and in opposite directions within one fine-tuning run: early on it credits boxes in reasoning that never finished (at temperature 1 it reports 24.1% on AIME 2025 when only 0.9% of responses deliver a correct answer), and later it misses correct answers stated without a box. Scoring instead the answer each finished response delivers (*delivered accuracy*) changes four experimental conclusions: a pre-registered fix that raised strict AIME accuracy after an empty thought by 31.0 percentage points did not improve delivered accuracy. On GPQA, strict grading also puts four of twelve public reasoning models at 0.0–11.2% while their delivered accuracy is 35.6–63.6%, and reorders 13 of 66 model pairs. An exact identity, valid for any model, writes the gap as credit for boxes in unfinished reasoning minus correct answers left unboxed, so two counts that need no judge bound it. The bound holds in 237 of 238 cells across three students and twelve public models, the larger count gives the error's sign in 99 of 110 cells where both are nonzero, and removing only the boxes from otherwise identical targets reproduces the late error in a second model (two seeds). Below a combined 2 points, the counts certify a strict score to within about 2 points without a judge; above it, we give a reporting protocol.

1 INTRODUCTION

Reasoning models are fine-tuned, selected and compared by the accuracy of their final boxed answers. Default graders in NeMo-Skills, NeMo-RL and verl read the last `\boxed{}` in a generation, even when the reasoning never finished, and credit no unboxed answer. The rule already misleads on released models. On GPQA it puts four of twelve public reasoning models at 0.0–11.2% while 35.6–63.6% of their responses deliver a correct answer, and it reorders 13 of 66 model pairs, 6 robustly (§7). We show that during supervised fine-tuning (SFT) it also misreads progress in both directions within a single run. When Qwen3-30B-A3B-Thinking is fine-tuned on Gemini-written solutions and sampled at temperature $T{=}1$, the grader reports 24.1% on AIME at step 25 although 95.7% of AIME generations hit the token budget and only 1 of 116 (0.9%) delivers an answer; by step 200 the student writes many answers in prose, and the grader misses them (Figure 1).

We compare the grader with *delivered accuracy*: whether the final answer that a finished response states in its visible channel, in any form, is correct. A judge that does not see the gold answer reads it (Qwen3-30B-A3B-Instruct-2507, checked against judges of two other families; §4). A response that hits the budget or never closes its reasoning delivers nothing. A strict score, by contrast, mixes delivered correctness with termination and answer format, in proportions that change during training. In practice, choosing checkpoints by strict accuracy costs 7.4 points of delivered accuracy on held-out decoding seeds, and a judge-free reader that respects termination cuts the loss to 0.6 (§7).

A strict score is the sum of three terms (§4): correct boxes after finished reasoning, in reasoning that hit the budget, and in reasoning that never closed; an answer stated without a box scores nothing. Two counts, available before any judge, therefore bound the gap between strict and delivered accuracy, up to rare parser–judge disagreements: unfinished responses that contain a box can only inflate the score, and finished responses without one can only deflate it. Both come from finish reasons, the thought closer and the parser, so existing evaluation logs can be audited without a judge.

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.

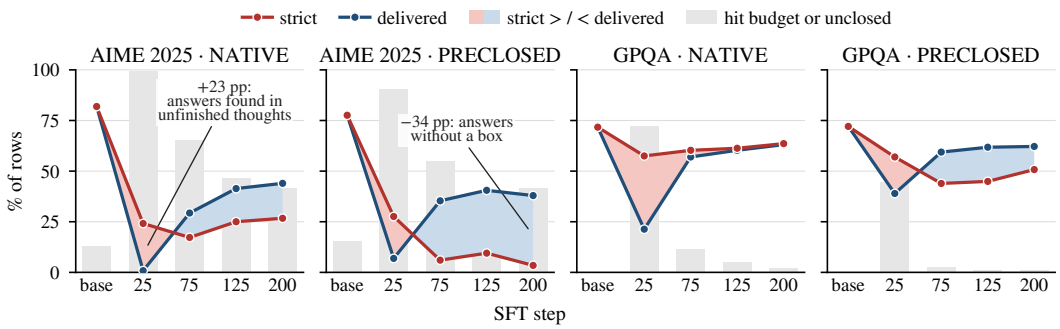

Figure 1: **In one SFT run, the strict grader overstates delivered accuracy early and understates it later.** Red line: strict (last-`\boxed{}`) accuracy; blue line: delivered accuracy, the answer a finished response delivers in its visible channel; red fill: strict above delivered (boxes found in unfinished thoughts); blue fill: strict below delivered (answers without a box); grey bars: rows that hit the budget or ended with the thought open. Qwen3-30B-A3B-Thinking fine-tuned on the Gemini recipe (training seed 42), $T{=}1$; base: before SFT. Prompt contracts: NATIVE, generation starts inside the thought; PRECLOSED, after an appended empty thought. AIME 2025: 29 items; GPQA: GPQA Diamond, 197 items; pp: percentage points.

SFT moves the two on different clocks. Early, termination collapses, yet 42–61% of the boxes in unfinished step-25 reasoning are correct, against chance rates of 0.1% on AIME and 25% on GPQA: answers the model found but never delivered. Later, termination recovers and the student adopts its teacher's answer format. Prompted after an appended empty thought, the student distilled from a prose-style teacher leaves 16% of its finished GPQA answers and 82% of its finished AIME answers unboxed at step 200.

We met the problem while comparing two prompt *contracts*, generation inside the thought (NATIVE) and after an appended empty thought (PRECLOSED), for students distilled from Gemini's prose-style solutions (the *Gemini recipe*) and gpt-oss-120b's boxed solutions (the *gpt-oss recipe*). Strict grading said that pre-closing hurt the Gemini student at step 200. A pre-registered fix appended gold boxed answers to 99.8% of its training targets and passed its registered test: the strict AIME cost of pre-closing shrank by 17.2 points, and strict pre-closed AIME accuracy rose by 31.0, yet delivered pre-closed accuracy changed by -3.5 points (its 95% CI rules out a gain of 5 or more). It passed every safeguard we had (pre-registration, fresh seeds, a manipulation check), because each tested the strict score, not what it measures (§8). Re-scoring the same outputs changes four decisions: the step-200 cost of pre-closing, the step-25 tie between the contracts, the gold-box rescue and a dip in the learning curve (§5).

We make four contributions. (i) **Theory.** An exact identity, valid for any model, writes the gap between strict and delivered accuracy as credit for boxes in unfinished reasoning minus correct answers left unboxed, and bounds it by two judge-free counts (§4). (ii) **Analysis.** Extraction errors are known to vary across models, prompts and budgets (Molfese et al., 2025; Jo et al., 2025; Hochlehnert et al., 2025); SFT moves the two terms on different clocks, so within one run the same grader first overstates and then understates delivered accuracy (§6). (iii) **Evidence of generality.** The bound holds in 237 of 238 cells across three students and twelve public models; where both counts are nonzero, the larger gives the gap's sign in 99 of 110, and in 27 of the 28 registered cells; removing only the boxes from otherwise identical targets causes the late underestimate in Qwen3-4B-Base, a model that SFT improves (§7, Appendix N). (iv) **Guidelines.** A judge-free check of when a strict score can be trusted, what to report instead, and why the usual safeguards missed the problem (§8).

2 Related Work

Answer-extraction choices change conclusions (Molfese et al., 2025; Hochlehnert et al., 2025; Jo et al., 2025), much prompt sensitivity disappears under an LLM judge (Hua et al., 2025), and trained or symbolic verifiers (Chen et al., 2025a; Liu et al., 2025b) improve how answers are compared rather than which text is read. Closest to ours, rule-based verifiers in reinforcement learning miss more correct answers as policies strengthen (Huang et al., 2025; Xu et al., 2025), an error in

one direction where ours changes sign; across long-CoT SFT runs termination tracks accuracy, attributed to unterminated outputs yielding no answer (Kopiczko et al., 2026), whereas the strict grader *credits* many of them. Completed and stopped-run answers are best reported apart (Carlon et al., 2026), and our identity says when this matters. On budgets, prompting a distilled reasoner to skip thinking with a dummy thought is competitive (Ma et al., 2025) (our PRECLOSED arm), and budgets change rankings (Sun et al., 2025; Jurayj et al., 2025; Ghosal et al., 2025). Checkpoint studies report accuracy dips with length surges during long-CoT SFT (Ren et al., 2026; Luo et al., 2025) and rank data recipes by strict scores (Guha et al., 2025; Yeo et al., 2025); our identity shows how much of such a dip the grader can make.

3 SETUP

The main program applies full-parameter SFT, on 100,000 mathematics questions from NVIDIA's Nemotron-Math-v2 (Du et al., 2025), to Qwen3-30B-A3B-Thinking-2507 (Yang et al., 2025a), a strong thinking model that scores 81.9% on AIME 2025 at our budget. The *Gemini recipe* uses targets we generated with `gemini-3-flash-preview` and a thought-prefill prompt, so each target contains a written thought and an answer; 0.08% contain a box. The *gpt-oss recipe* uses the corpus's own gpt-oss-120b solutions (100% boxed). Both recipes close the thought on 100% of targets. The optimizer, learning rate 10^{-5}, seeds and schedule are identical, with checkpoints every 25 steps. The recipes differ in teacher and content as well as answer format, so their comparison does not isolate format; §6 isolates it.

Evaluation uses AIME 2025 and GPQA Diamond (Rein et al., 2024). We exclude each task's piloted first item, leaving 29 and 197 items, respectively, and use four decoding seeds with $T{=}1$ and top-p 1, the sampling distribution of pass@k and RL rollouts. The 32,768-token budget matches the training sequence length. The prompt requests an answer in `\boxed{}` on AIME and a final line `Answer: \boxed{A/B/C/D}` on GPQA.

The registered families test the contract *interaction*, defined as PRECLOSED$-$NATIVE at step 200 minus the same contrast at step 25, on the training seed and two fresh seeds, together with two *box-appending retrains* that add either a heuristic boxed answer to 62% of the Gemini targets or a gold boxed answer to 99.8%. The retrains test the change ΔR in contract reliance $R =$ NATIVE$-$PRECLOSED at step 200. Each family's code and thresholds were frozen and audited before any outcome was read, and every result was recomputed independently. Every delivered-accuracy analysis of this program is post hoc; the held-out and public-model tests (§7) were registered in advance. We use item-clustered bootstrap intervals and permutation tests (10,000 resamples), treat training seeds as fixed replicates, and apply Holm correction within each registered family. *Significant* means that the 95% CI excludes zero; every headline contrast that the text calls significant also survives a max-$|t|$ adjustment over the 14 headline contrasts of Table 10 (bold in tables and figures is unadjusted). Non-significant contrasts are not interpreted as evidence of no effect.

The Gemini recipe lowers the strong base model's delivered AIME accuracy from 81.9% to 44.0% at step 200, while the gpt-oss recipe lowers it less, to 67.2%. The termination collapse recurs at each tested learning rate, later at lower ones, on every training seed and in both distilled R1 students (§7), and never in the base model. *Gemini* and *gpt-oss* name the teachers of the targets; in the main program the student is always the same model. Distillation from another model's traces is standard (Guo et al., 2025; Guha et al., 2025); our student starts strong.

4 A STRICT SCORE MIXES DELIVERY, FORMAT AND CORRECTNESS

Strict scoring reads the last `\boxed{}` after the last thought closer, or anywhere in the output if it has no closer: it credits boxes in unfinished thoughts but never unboxed answers. *Delivered* accuracy scores the answer a finished response states in its visible channel, boxed or not; unfinished rows deliver nothing (§1). The *gap* is strict minus delivered accuracy (the appendices say *committed* accuracy and *artifact*); a *cell* is one checkpoint, task and contract.

Every response ends in one of three *termination states* (Figure 2): it hit the budget (B), emitted end-of-turn with its thought channel still open (U, only under NATIVE), or closed the channel and

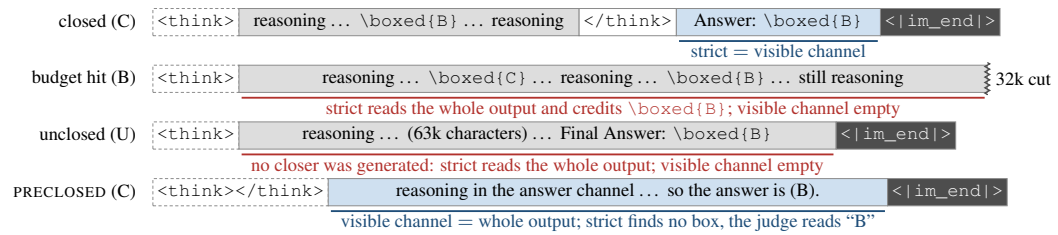

Figure 2: **The strict rule reads inside unfinished thoughts, the judge only the delivered answer.** Termination states (B: budget hit; U: unclosed; C: closed) and scorer inputs. Rows 1–3: NATIVE (generation starts inside thought); row 4: PRECLOSED (after empty thought). Grey: thought; blue: visible channel; dashed: prompt; dark: end-of-turn; 32k cut: budget. Underline: what the strict (last-`\boxed{}`) rule reads, red if inside unfinished thought; the judge reads only the visible channel.

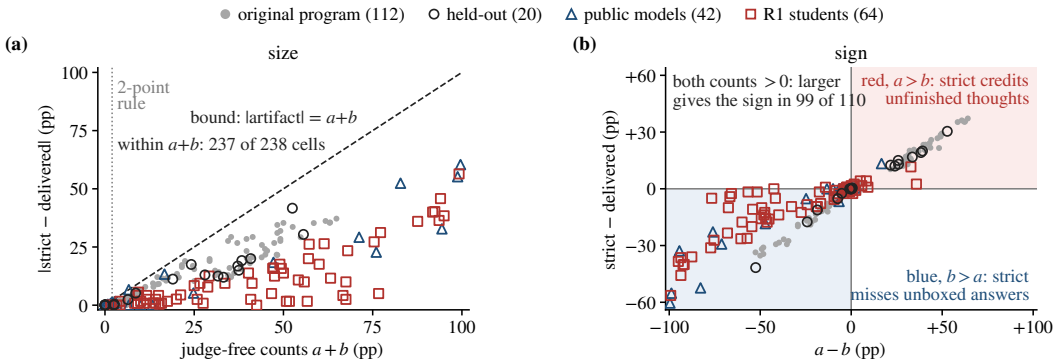

Figure 3: **Two counts that need no judge bound the gap by construction and, in most cells, predict its sign.** a: share of rows that are unfinished but contain a box; b: share of rows that are finished but contain none. (a) $|\text{strict} - \text{delivered}|$ against $a{+}b$; dashed: $y{=}x$, which equation 3 guarantees up to the disagreement term δ; dotted: $a{+}b = 2$ pp, free screen. (b) $\text{strict} - \text{delivered}$ against $a - b$; unless one count is zero the identity does not force the sign, so the shaded quadrants (larger count gives the sign) are a finding, weakest where counts are large. Cells: original program; held-out items; twelve public models (the 42 registered cells); the two distilled R1 students. The Gemma and Qwen3-4B cells are reported in the appendices. pp: percentage points.

ended (C). Let π be the event that the strict rule finds a well-formed `\boxed{}` answer in the text it reads (after the last closer if there is one, else the whole output), y_π that answer, y_c the answer a closed response commits to in its visible channel, in any form, and g the gold answer. Then

$$\text{Strict} = \underbrace{P(\text{C}, \pi, y_\pi{=}g)}_{\text{delivered and boxed}} + \underbrace{P(\text{B}, \pi, y_\pi{=}g) + P(\text{U}, \pi, y_\pi{=}g)}_{\text{found in unfinished thoughts}}, \tag{1}$$

$$\text{Delivered} = P(\text{C}, \pi, y_c{=}g) + \underbrace{P(\text{C}, \neg\pi, y_c{=}g)}_{\text{delivered without a box}}. \tag{2}$$

Subtracting gives bounds that hold by construction (the exact identity is equation 4, §6),

$$-\underbrace{P(\text{C}, \neg\pi)}_{b} - \delta \;\le\; \text{Strict} - \text{Delivered} \;\le\; \underbrace{P(\pi,\, \text{B} \cup \text{U})}_{a} + \delta, \qquad \delta = P(\text{C}, \pi, y_\pi{\neq}y_c), \tag{3}$$

so $|\text{gap}| \le \max(a, b) + \delta$, where a is the share of rows that are unfinished (B or U) and boxed and b the share that are finished (C) and unboxed; we registered and plot the looser $a{+}b$. Parser–judge disagreement is rare ($\delta = 0.21\%$ of rows), so the two judge-free counts carry the diagnostic.

In all 112 cells of the original program (28 checkpoints × 2 tasks × 2 contracts), $|\text{gap}| \le a{+}b$, with the gap typically about half of $a{+}b$. The larger count gives the sign of 91 of the 97 nonzero gaps, and of 59 of the 63 with both counts nonzero, where the identity does not force it (Figure 3b). When $a{+}b < 2$ points, the gap is at most 1.14, so the counts give a free screen for when a strict score can be trusted. These cells share models, items and neighbouring checkpoints, so the counts are not independent replications.

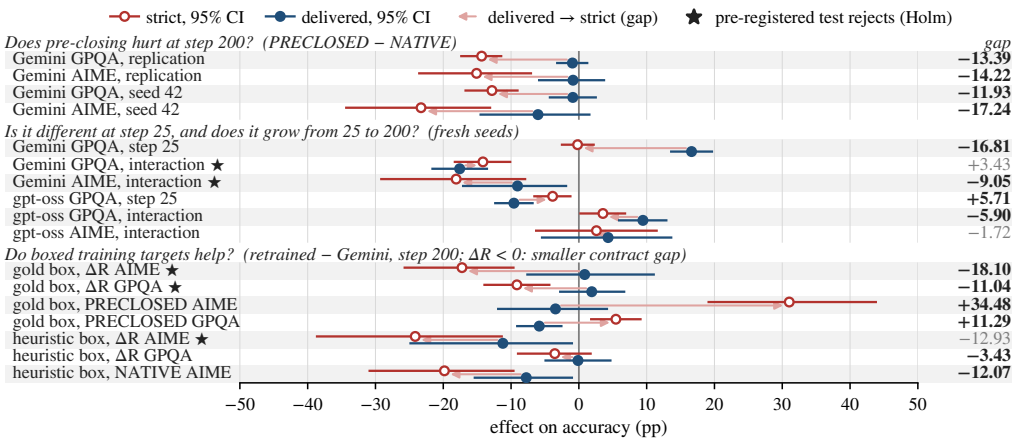

Figure 4: **Most of the program's large strict effects shrink under delivered-accuracy scoring, and some reverse.** Strict (last-`\boxed{}`; red) and delivered (judge-read visible answer; blue) estimates with item-cluster 95% CIs; right column: gap (strict−delivered), bold if its CI excludes 0; $\star$: pre-registered strict test rejected (Holm-adjusted permutation $p < .05$). Contrasts are PRECLOSED−NATIVE (generation after minus inside the thought) unless stated; interaction: step 200 minus step 25; replication: mean of two fresh training seeds; ΔR: change in contract reliance R = NATIVE−PRECLOSED after a box-appending retrain. AIME: AIME 2025 (29 items); GPQA: GPQA Diamond (197 items); Gemini, gpt-oss: recipes named for their teachers.

For delivered accuracy, an instruction-tuned judge (Qwen3-30B-A3B-Instruct-2507) reads the last 3,000 characters of the visible answer channel and returns the single final answer the response commits to, or NONE; it sees neither the gold answer, the strict parse nor the problem. A *robust boxed* rule (balanced braces, stripped wrappers) moves no contract contrast by more than 2.6 pp, since a better comparator cannot credit an undelivered answer or read one the parser never extracts.

The primary judge shares a model family with the student, so we validate it with judges from other families. A second judge, gemma-3-12b-it, agrees on 99.44% of 40,384 terminated rows and moves no delivered-accuracy contrast by more than 0.9 pp; a third, gpt-oss-120b, keeps every gap's sign. The judged answer occurs literally in the judged tail for 99.80% of answers, and verdicts survive 99.97% of 26,785 answer-preserving rewrites. Counting every hedged answer as undelivered leaves the sign of every reported gap unchanged.

5 RE-READING THE ANSWERS CHANGED FOUR CONCLUSIONS

Figure 4 shows the 14 headline contrasts, and three related ones, under both scorers, and Table 1 re-reads each decision seven ways (all post hoc). The step-25 tie depends on crediting boxes in unfinished thoughts: readers that remove this credit put PRECLOSED ahead. The step-200 cost of pre-closing and the apparent gold-box rescue depend on how finished answers are read: judging them removes both conclusions, with or without that credit. Counting hedged answers as undelivered leaves a GPQA cost of 4.4 points, and a lenient regex keeps both on AIME, where students write integers in prose ("Result is 167."). The step-25-to-75 strict dip depends on both.

At step 200, the apparent cost of pre-closing is mostly a grading artifact. On the fresh seeds, the Gemini GPQA contrast is -14.34 pp under strict scoring but -0.95 pp under delivered scoring; on AIME, the corresponding estimates are -15.09 and -0.86. Both gaps are significant, and the seed-42 run agrees. This late gap does not mean that the registered interaction disappears. At step 25, the NATIVE arm ends inside an unclosed thought on 21–24% of GPQA rows, while under PRECLOSED the whole output is visible. Delivered scoring therefore puts PRECLOSED ahead by 16.62 pp where strict scoring sees a tie (-0.19). On GPQA the interaction is even larger under delivered scoring (-17.58 versus -14.15 pp strict; on AIME it halves, -9.05 versus -18.10), but it reflects an early failure that later disappears rather than a late cost. Crediting unfinished boxes while judging finished answers removes both sources and leaves no clear interaction (-3.4 on GPQA and -5.2 pp on AIME, neither statistically significant).

Table 1: **Seven readings of the same generations change experimental decisions.** Original items, fresh training seeds except seed 42 for the rescue and dip. Strict: registered rule; reasoning parser: strict with budget-hit and unclosed rows unanswered; lenient regex: last box, else an "Answer:" or "answer is" line, in the visible channel; delivered: judge-read delivered answer; + unclosed credited: unclosed rows keep their judged answer; hedged = none: hedged answers are undelivered; + unfinished boxes: delivered plus strict credit for boxes in budget-hit and unclosed rows. Entries (pp): PRECLOSED−NATIVE (generation after minus inside the thought) unless marked; ΔR: change in reliance R =NATIVE−PRECLOSED (negative: rescue); Δ acc.: step 75 minus step 25 accuracy. Bold: unadjusted 95% CI excludes 0; shaded: verdict differs from strict.

		judge-free readers			readers using a judge			
decision	task	strict	reasoning parser	lenient regex	delivered	+ unclosed credited	hedged = none	+ unfinished boxes
pre-closing hurts, step 200	GPQA	**−14.3**	**−13.6**	**−2.6**	−0.9	−2.1	**−4.4**	−1.7
	AIME	**−15.1**	**−15.1**	**−11.2**	−0.9	−0.9	−0.9	−0.9
the effect grows, 25→200	GPQA	**−14.2**	**−28.3**	**−18.9**	**−17.6**	**−5.3**	**−20.1**	−3.4
	AIME	**−18.1**	**−22.0**	**−18.5**	**−9.1**	−3.0	**−9.1**	−5.2
better contract, step 25	GPQA	−0.2	**+14.7**	**+16.2**	**+16.6**	**+3.2**	**+15.7**	+1.7
gpt-oss: early cost	GPQA	**−3.9**	**−9.8**	**−9.4**	**−9.6**	**−9.9**	**−9.9**	**−3.6**
gold boxes rescue (ΔR)	AIME	**−17.2**	**−18.1**	**−12.1**	+0.9	+1.7	0.0	+1.7
	GPQA	**−9.1**	**−9.5**	+0.1	+1.9	+2.3	−2.3	+2.3
dip, steps 25→75 (Δ acc.)	AIME	**−6.9**	**+16.4**	**+16.4**	**+28.4**	**+25.9**	**+28.4**	**+5.2**

The box-appending retrain shows the same split between format and delivered correctness. Appending gold boxes raises the strict parse rate on PRECLOSED AIME from 10 to 52% without raising delivered accuracy. The change in contract reliance, ΔR, is -17.24 on AIME and -9.14 on GPQA under strict scoring, but $+0.86$ and $+1.90$ pp under delivered scoring. The apparent rescue thus comes mostly from boxing rather than more correct answers: the delivered CIs exclude a rescue larger than 7.8 points (AIME) or 2.9 (GPQA). On pre-closed GPQA the retrain even lowers delivered accuracy (-5.84 pp) while strict reports $+5.46$.

The learning curve is also reinterpreted. The strict NATIVE AIME curve dips ($24.1 \to 17.2 \to 25.0 \to 26.7\%$), while delivered accuracy rises ($0.9 \to 29.3 \to 41.4 \to 44.0\%$; Figure 1). As termination recovers, credit for unfinished thoughts vanishes and prose answers appear. On the two fresh seeds, decoded under one engine configuration, strict accuracy falls by 9.1 points from step 25 to 75 (95% CI -17.7 to -0.9; -15.5 and -2.6 by seed) while delivered accuracy rises by 31.5 points. The rise is robust; the depth of the dip varies with the seed and with the engine configuration, which alone moves step-25 strict accuracy by 4–6 points (Appendix F.5).

The gpt-oss student's early cost of pre-closing, unlike the Gemini student's late one, keeps its sign under every reading, but its size depends on how truncated answers are counted. At step 25, its PRECLOSED arm hits the budget on 50–53% of GPQA rows, compared with 35–39% under NATIVE. An answer is stated before the cut on 47.5% of its PRECLOSED rows, usually restated in a loop (median 18 statements); 43% of these rows state two or more different answers. The cost is -9.6 pp when truncated rows deliver nothing, -5.1 when answers never revised are credited, and -1.5 when every stated answer is credited (not significant); strict scoring reports -3.9.

Re-scoring preserves the termination collapse. At step 200 the Gemini contract estimates are small: -6.0 to $+2.6$ pp on AIME across three seeds and about -1 pp on GPQA, with delivered CIs (seed 42; fresh seeds pooled) that include zero; counting hedged answers as undelivered gives -4.4 to -5.1 pp on GPQA.

6 EARLY COLLAPSE AND LATE FORMAT DRIFT PRODUCE OPPOSITE ERRORS

The early error accompanies a termination collapse that extends beyond the lengths seen during training (Figure 5a). Before over-length targets are dropped, teacher thoughts have medians of 12,221 tokens for Gemini and 11,448 for gpt-oss; no trained target exceeds the 32k budget, and the base model's GPQA outputs never reach it. After 25 SFT steps on either recipe, however, the median GPQA output is 19–32k tokens, and 35–53% of GPQA rows and 77–96% of AIME rows hit the budget. Across the step-25 students, 29–87% of budget-hit rows contain boxes that the strict rule can credit, compared with none of the base model's 33 budget-hit rows (all on AIME). A

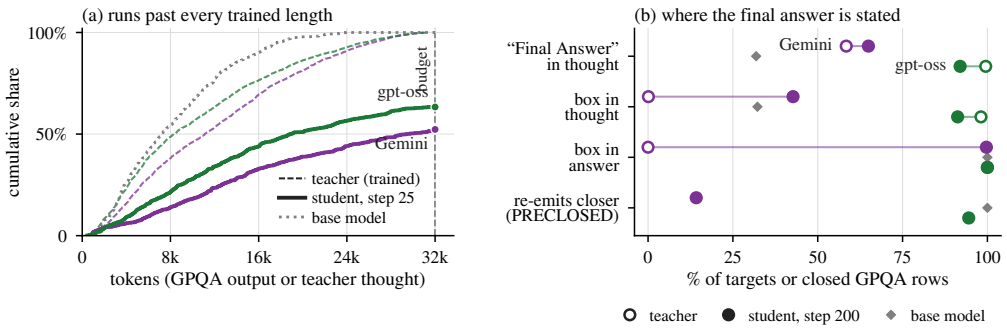

Figure 5: **Students first overrun every length they were trained on; later, the Gemini student keeps its teacher's "Final Answer" habit and, under pre-closing, rarely re-closes its thought.** (a) Cumulative share of rows by length in tokens: the teachers' thoughts (a 3,000-target sample, targets kept for training, i.e. at most 32,768 tokens) and the GPQA outputs of the base model and of both step-25 students (NATIVE); dashed: the 32,768-token budget. (b) Where the final answer is stated (%): a "Final Answer" phrase or a `\boxed{}` inside the thought, a `\boxed{}` in the answer segment, and, under PRECLOSED (generation after an empty thought), a rollout that re-emits its own thought closer; teachers: 3,000 training targets; students (step 200) and base model: closed GPQA rows, NATIVE except the last row.

self-distilled student, which sees no teacher shift, does not collapse at step 25 (exploratory; one seed, another harness): it hits the budget on 2.5% of AIME rows and scores 81.7% under both scorers.

The later error reflects where and how the student states its answer (Figure 5b). The Gemini teacher writes "Final Answer" inside its thought on 58% of targets and boxes 0.1%, whereas the gpt-oss teacher boxes in both channels. Under NATIVE, both students close the thought and box on GPQA, like the base model. Under PRECLOSED, the base model and the gpt-oss student treat the empty thought as an instruction to think anyway: they reason in the answer channel, re-emit the closer and box. The Gemini student instead re-emits the closer on only 14% of GPQA rows and otherwise leaves its derivation in the answer channel, ending in prose on AIME.

A controlled test isolates the format. We trained Qwen3-4B-Base on the gpt-oss targets and, in a registered twin, on the same targets with every `\boxed{}` replaced by its content; SFT improves this model (unregistered: PRECLOSED delivered accuracy 19.2 vs 5.0% on AIME 2024 and 34.4 vs 32.2% on GPQA-main, the GPQA items outside Diamond, step 200 vs base). At step 200 under PRECLOSED the unboxed twin scores 0.0 and 6.0% strict but 18.3 and 27.6% delivered on AIME 2024 and GPQA-main, gaps of -18.3 and -21.6 points; the boxed student's are $+1.7$ and -3.8. Strict accuracy separates the twins by 21–25 points, delivered accuracy by at most 7; a second, pre-registered seed gives 17–20 and 4 and passes all three of its registered predictions; one of the first seed's failed, as its boxed student's GPQA gap fell outside ± 3 points (Appendix N).

Early collapse and late format drift contribute opposite terms to the gap. Subtracting equation 2 from equation 1 gives

$$\text{Strict} - \text{Delivered} = \underbrace{P(\text{B}\cup\text{U}, \pi, y_\pi{=}g)}_{\text{credit (early)}} - \underbrace{P(\text{C}, \neg\pi, y_c{=}g)}_{\text{unread (late)}} + \epsilon, \qquad |\epsilon| \le \delta, \tag{4}$$

where ϵ is the parser–judge disagreement among closed, parsed rows. The credit term peaks with the collapse at step 25; the unread term appears once the student adopts its teacher's answer form in the channel the contract exposes. At $T{=}1$, a strict curve therefore combines a decaying positive term with a growing negative one. It can dip where they cross and shows a "contract effect" wherever the negative term differs between arms.

7 THE GAP RECURS ACROSS MODELS, AND TWO COUNTS PREDICT ITS SIGN

We separate transfer of the gap from transfer of its sign (Figure 6, Table 2). The held-out panel (items never evaluated in the program: AIME 2024 and "GPQA-main", the 250 GPQA items outside

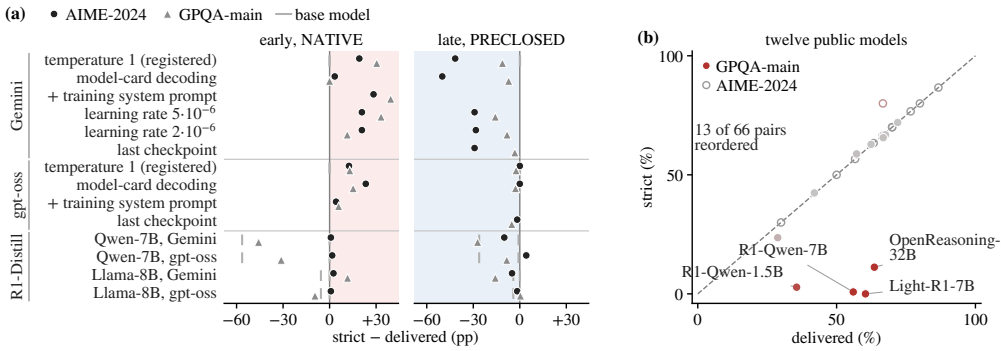

Figure 6: **Gaps recur across training settings and models, and their sign follows how each model states its answers.** (a) Gap (strict−delivered, pp) on the held-out items (AIME 2024, 30 items × 4 seeds; GPQA-main, 250 items × 2 seeds): early, the step-25 NATIVE cell (lower learning rates, where accuracy bottoms out: step 50 at $5{\cdot}10^{-6}$, 100 at $2{\cdot}10^{-6}$); late, the step-200 PRECLOSED cell ($5{\cdot}10^{-6}$: 200; $2{\cdot}10^{-6}$: 375; last checkpoint: Gemini 885, gpt-oss 702); rows: our Qwen3 student under each setting, and the two R1 students; circles: AIME; triangles: GPQA; grey ticks: base model; red and blue tints: credit for unfinished thoughts, unread answers. (b) Twelve public reasoning models on GPQA-main (filled) and AIME 2024 (open), NATIVE, recommended decoding: strict against delivered accuracy; points below the diagonal: strict undercounts; named: the four models whose strict and delivered accuracies differ most; NATIVE/PRECLOSED: generation inside/after an empty thought.

Diamond) carries the registered predictions of Table 2. Exploratory runs on it show the gap at lower learning rates, with the training-time system prompt and at the last checkpoint, where strict still understates the Gemini student on AIME 2024 by 21.7 (NATIVE) and 29.2 points (PRECLOSED). Strict grading also reorders 13 of 66 pairs of twelve public reasoning models on GPQA-main (6 remain significant after multiple-comparison adjustment). Every reordered pair involves one of five models that leave 25–100% of finished rows unboxed, which a large b (finished rows without a box) flags; Table 1's lenient regex reorders none.

An exploratory Gemma-4-31B-it panel (60 cells, not a matched design) shows the same gap: for gpt-oss-recipe checkpoints, strict is 15–38 points below delivered accuracy on GPQA in both contracts (Appendices R, H). The judge-free bound holds in 237 of 238 cells of Figure 3. Where one count is zero the identity fixes the sign up to parser–judge disagreement δ; where both are nonzero, the larger gives it in 99 of 110 cells, and in 35 of 43 where both reach 2 points; each miss has $|\text{gap}| \leq 2.5$ points.

Transfer of the gap does not imply a universal temporal reversal, which needs both early credit for unfinished boxes and later unboxed answers. Of the Qwen3 student's 50 trajectories (a recipe, run, task and contract across checkpoints), 27 move from a significantly positive to a significantly negative gap, and none the reverse; early credit nearly vanishes under greedy and model-card decoding for the Gemini student, whose loops hold no box, but not for the gpt-oss student. DeepSeek-R1-Distill-Qwen-7B (R1-Qwen), trained on both recipes with predictions frozen before any checkpoint existed, reproduces the step-25 collapse but states GPQA letters without a box from the start (base: 0.6% strict, 57.0% delivered), so its GPQA gap is negative throughout, as $b > a$ predicts; none of its eight trajectories is ever significantly positive, and its registered directional predictions fail (Appendix M.1). DeepSeek-R1-Distill-Llama-8B (R1-Llama), registered the same way, shows both halves for the Gemini recipe: its step-25 NATIVE gaps are positive on both tasks, and every step-200 cell whose count b exceeds a by 5 points has a negative gap (-4.2 to -15.8; addendum frozen before any judgement; Appendix M.2). Its contract-level predictions fail because strict grading misses answers under both contracts alike.

Strict and delivered accuracy pick different checkpoints in 23 of 58 fine-tuned trajectories (Qwen3 and R1-Qwen). Picked on half the decoding seeds and scored on the rest, strict's choice loses 7.4 points of delivered accuracy on average over the 56 with split seeds; a judge-free reader (no answer for unfinished rows; the box, else an answer line) loses 0.6.

Table 2: **Prospective tests separate recurring grading error from student-specific dynamics.** ✓: holds; ✗: fails; —: not registered. New items: held-out panel, P1–P9 frozen before generation; public: twelve models, predictions frozen before any output; R1-Qwen, R1-Llama: S1–S6 frozen before any checkpoint existed; A1–A3: R1-Llama addendum, written after the R1-Qwen outcomes and frozen before any judgement was read. The first two rows hold by construction up to δ, as does the sign row wherever one count is zero (all public cells; 8 of the 20 A1 cells). †: registered with $+\delta$; the judge-free form is shown. ‡: S4 bounds the difference between the contracts' gaps, P5 each cell. Counts: cells meeting the prediction / cells where it applies.

prediction	new items	public	R1-Qwen	R1-Llama
identity, up to δ				
judge-free bound $\lvert\text{gap}\rvert \le a{+}b$	P1 ✗ 19/20 (one row)	✓ 42/42	S5 ✓ 36/36†	S5 ✓ 28/28†
2-point rule: $a{+}b < 2 \Rightarrow \lvert\text{gap}\rvert \le 2\,\text{pp}$	P1 ✓ 5/5	✓ 16/16	S6 ✓ 1/1	S6 ✓ 1/1
empirical				
larger of a, b gives the sign	—	✓ 14/14	post hoc 29/34	A1 ✓ 20/20
step-25 collapse (budget-hit, unclosed)	P2 ✗ (misses 1 of 9 thresholds)	—	S1 ✓	S1 ✓
step-25 credit: strict $>$ delivered	P3 ✓ 8/8	—	S2 no eligible cell	S2 ✓ 4/4; A2 ✓
step-200 Gemini: strict understates PRECLOSED more	P4 ✓	—	S3 ✗ opposite sign	S3 ✗ both contracts
step-200 gap < 0 where $b - a \ge 5$	—	—	—	A3 ✓ 4/4
gpt-oss step-200 gap within ± 3 pp‡	P5 ✗	—	S4 ✗	S4 ✗
strict recipe gap exceeds the delivered gap	P7 ✓	—	—	—
base: no gap; state orders correctness; hedging	P6 ✓; P8 ✓ (pooled); P9 ✓	—	—	—

8 WHY STANDARD SAFEGUARDS MISS THE ERROR, AND WHAT TO REPORT

All the program's checks passed, yet none tested whether the strict score measured delivered correctness: pre-registration fixed a strict contrast without validating it, fresh seeds reproduced a recipe property, termination looked normal at step 200 while the loss lay in closed rows, and the gold-box retrain's manipulation check passed because boxing, which the strict grader rewards, rose.

Guidelines. (1) Report separately, for every checkpoint and contract, how often responses finish, state an answer and state a correct one, and never credit a box in an unfinished thought. (2) Before any judge, count a (unfinished rows with a box) and b (finished rows without one): below $a{+}b = 2$ points the bound caps the gap, and here the gap never exceeded 1.14. (3) Above it, read the delivered answer with a validated judge (18 GPU-minutes per 40k rows here); a judge-free reader (unfinished: no answer; else the box or an answer line) selects checkpoints nearly as well as delivered accuracy and reads letter answers, but misses integers written in prose (off by over 2 points in 26 of the 112 original cells; strict: 71). (4) Training targets should state the final answer in the visible channel; an appended box should agree with the reasoning.

9 LIMITATIONS

Delivered accuracy uses a validated judge, not human labels; contested rows had only a non-blind 100-row review by an author-supervised AI agent, and the judge may be lenient on hedged prose. Judges were validated only on integer and letter answers; the visible-channel rule was fixed post hoc, after a whole-output reading (Appendix E). Only Qwen3 is seed-replicated, and only answer format is isolated causally. Early credit depends on decoding, AIME has 29 items, and scorers that give unfinished rows no credit penalise models needing larger budgets.

10 CONCLUSION

Strict `\boxed{}` grading confounds delivered correctness with termination and answer format, and fine-tuning shifts the mix; separating them removed a spurious cost, repair and dip from our conclusions and revealed a hidden early advantage. For any model, two judge-free counts bound the error, so existing logs can be audited; reporting termination, delivery and correctness separately avoids it.

ETHICS STATEMENT

This work evaluates language models on public benchmark items and trains on NVIDIA's public Nemotron-Math-v2 and Llama-Nemotron corpora (CC BY 4.0 and CC BY-SA 4.0, attributed in Appendix B) and on targets we generated for the same mathematics questions; no human subjects, personal data or crowd workers are involved. We do not redistribute AIME problem statements; the released artifact binds items by hash to a copy the reproducer supplies lawfully, and GPQA is used under its license. The paper's societal implication is methodological: widely used evaluation practice can misattribute capability, which we aim to make visible. We report the program's initial interpretation, and why the evidence overturned it, in full.

REPRODUCIBILITY STATEMENT

Every pre-registered family of the program (§3) was frozen before any response existed, with an independent design audit, an outcome-blind completion gate, atomic shard publication, a hashed analyzer and an independently implemented exact recomputation (Appendix D). The held-out test froze its protocol, panel and predictions before generation; its analyzer was hash-locked while generation was running, amended once and re-locked, all before any response was read, and then run once (Appendix J). The strict rule, the robust-boxed and delivered-accuracy scorers, the judge prompt, decoding settings and validation statistics are given in §4 and Appendices B, C and E. The supplementary artifact contains the per-row judge outputs (answers only; both judge families for the original program and the held-out panel, the first judge alone for the second evaluation harness of Appendix O); the judge prompt; per-cell tables under the three scorers for the original program, labelled by commitment policy, and the frozen JSON behind every figure and table; an official-source hash binding for the 198 GPQA items and a hash-binding interface for the 30 AIME items; the held-out pre-registration, amendment, code locks, analyzer, results and second-judge comparison; the public-model pre-registration, receipt, code lock, panel manifest, per-row judge outputs and results; the pre-registrations, per-row judge outputs and results of the R1-Qwen, R1-Llama and Qwen3-4B-Base tests and of the fresh-seed dip; the result files and scripts of the exploratory checks (greedy, model-card and presence-penalty decoding, the 64k budget, model-card decoding at step 200, the every-5-steps rerun, the learning-rate checkpoints, the last checkpoints and the system-prompt variant on the held-out panel, a symbolic-equivalence checker, the Gemma-4-31B re-judge and the judge length panel) and the source file of every number bound to the text; a stand-alone script that recomputes every held-out cell from the released rows; a script that recomputes the paper's key numbers from the archive with the Python standard library, and a guide that maps every figure, table, appendix and registered prediction to its files; and the as-run analysis and figure scripts, hash-pinned. The row-level scripts read generations and benchmark text that are not redistributed, so replay starts from the released judge rows and JSON.

AI USE STATEMENT

Generative AI tools were used extensively in this work, under human direction and review. Required disclosures: they were used to propose and refine hypotheses (including the answer-sufficiency hypothesis, that SFT distils where the answer is stated, which this paper abandons); to design the experimental protocols, pre-registration documents, analysis locks and interpretation grids; to implement the evaluation harness, the training-data surgeries, the analyzers (including the held-out analyzer, hash-locked before any of its inputs was read and run once), the independent auditors and the committed-answer scoring; to support the interpretation of results; to perform, under author supervision, the 100-row reading of contested judge outputs reported in §9 and Appendix E (no human read these rows and no human labels were produced; the reader, itself a language model, worked from stored records holding the gold answer, both judges' answers and at most the last 320 characters of each response, so the reading was not blind, may share the judges' blind spots and is not an independent validation); and to help construct the surgical training sets by deterministic code (the gold answers come from the source data). Instruction-tuned language models are also used *as measurement instruments* (the committed-answer judges), which we validate in §4. They were not used to generate benchmark items. The identity, the two-sided bound and its proof, and the Shapley allocation of Appendix A.4 were drafted with AI assistance and checked by the authors and by an independent AI reviewer that recomputed every allocation; the teacher-generated distillation targets are the object of study. Recommended disclosures: literature search and summarisation,

figure generation, and drafting and editing this manuscript. We reviewed the AI-assisted work: all confirmatory statistics were recomputed by an independently implemented code path, the main-text table values and most other reported numbers are checked against frozen machine-readable sources by a separate script, and cited works were verified against their primary pages. We take responsibility for the final content of this work.

REFERENCES

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. arXiv:2511.04703.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, et al. Llama-Nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024. URL https://arxiv.org/abs/2405.14782.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2503.01747.

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

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

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.852. URL https://aclanthology.org/2024.findings-emnlp.852/.

Ding Chen, Qingchen Yu, Pengyuan Wang, Mengting Hu, et al. xVerify: Efficient answer verifier for reasoning model evaluations. *arXiv preprint arXiv:2504.10481*, 2025a.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT think that much for 2+3=? On the overthinking of long reasoning models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9487–9499. PMLR, 2025b. URL https://proceedings.mlr.press/v267/chen25bx.html. arXiv:2412.21187 (arXiv title: "Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs").

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.

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. Zenodo, version v0.4.0, doi:10.5281/zenodo.10256836, December 2023. URL https://zenodo.org/records/10256836.

Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, et al. Does thinking more always help? Mirage of test-time scaling in reasoning models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.04210.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.282. URL https://aclanthology.org/2025.findings-naacl.282/.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Mohsen Hariri, Weicong Chen, Nahal Shahini, Vikash Singh, Kai Ye, Amirhossein Samandar, Debargha Ganguly, Sreehari Sankar, Yanyan Zhang, Shouren Wang, Jerry Peng, Biyao Zhang, Michael Hinczewski, and Vipin Chaudhary. Test-time scaling in reasoning LLMs: Inference regimes, evaluation, and reproducibility. *arXiv preprint arXiv:2608.04001*, 2026. URL https://arxiv.org/abs/2608.04001.

Muyu He, Muhammad Ali Shafique, Anand Kumar, Tsach Mackey, and Nazneen Rajani. The valley of code reasoning: Scaling knowledge distillation of large language models. *arXiv preprint arXiv:2510.06101*, 2025. URL https://arxiv.org/abs/2510.06101. NeurIPS 2025 Workshop on Deep Learning for Code in the Agentic Era (DL4C).

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 17073–17114, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/18b3b2947174ae9b1a8fc91ff89f4eb6-Abstract-Conference.html. arXiv:2508.13144.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://arxiv.org/abs/2103.03874.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2504.07086.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

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. URL https://openreview.net/forum?id=rygGQyrFvH. arXiv:1904.09751.

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*, pp. 19889–19899, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1006. URL https://aclanthology.org/2025.emnlp-main.1006/.

Yuzhen Huang, Weihao Zeng, Xingshan Zeng, Qi Zhu, and Junxian He. From accuracy to robustness: A study of rule- and model-based verifiers in mathematical reasoning. *arXiv preprint arXiv:2505.22203*, 2025. URL https://arxiv.org/abs/2505.22203. v1 titled "Pitfalls of Rule- and Model-based Verifiers – A Case Study on Mathematical Reasoning".

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4615–4634, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.254. URL https://aclanthology.org/2024.acl-long.254/.

Hangzhan Jin, Sitao Luan, Tianwei Ni, Sicheng Lyu, et al. RL fine-tuning heals OOD forgetting in SFT. *arXiv preprint arXiv:2509.12235*, 2025.

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

William Jurayj, Jeffrey Cheng, and Benjamin Van Durme. Is that your final answer? Test-Time scaling improves selective question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 636–644, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-short.50. URL https://aclanthology.org/2025.acl-short.50/.

Dawid J. Kopiczko, Sagar Vaze, Tijmen Blankevoort, and Yuki M. Asano. Data repetition beats data scaling in long-CoT supervised fine-tuning. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2602.11149.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (SOSP)*, 2023. URL https://arxiv.org/abs/2309.06180.

Hynek Kydlíček. Math-Verify: Math verification library. Hugging Face, 2025. URL https://github.com/huggingface/math-verify.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.21140.

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

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. X-Reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025a. URL https://arxiv.org/abs/2505.03981.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, et al. CompassVerifier: A unified and robust verifier for LLMs evaluation and outcome reward. *arXiv preprint arXiv:2508.03686*, 2025b.

Yang Liu and Chenhui Chu. Understanding the prompt sensitivity. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44362–44388, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.2053. URL https://aclanthology.org/2026.acl-long.2053/.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11679–11698, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.584. URL https://aclanthology.org/2025.naacl-long.584/.

Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, et al. LLMs are biased towards output formats! Systematically evaluating and mitigating output format bias of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2408.08656.

Renjie Luo, Jiaxi Li, Chen Huang, and Wei Lu. Through the valley: Path to effective long CoT training for small language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4972–4992, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.251. URL https://aclanthology.org/2025.emnlp-main.251/.

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

Lovish Madaan, Aaditya K. Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024. URL https://arxiv.org/abs/2406.10229.

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

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2503.14996.

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

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. URL https://arxiv.org/abs/2512.12895.

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)*, 2024. URL https://arxiv.org/abs/2405.17202.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-Proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2311.12022.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, et al. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2604.06628.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 55565–55581, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html. arXiv:2304.15004.

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. URL https://arxiv.org/abs/2310.11324.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953.

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

Lukas Twist, Helen Yannakoudakis, and Jie M. Zhang. Reasoning-trace collapse: Evaluating the loss of explicit reasoning during fine-tuning. *arXiv preprint arXiv:2605.21127*, 2026. URL https://arxiv.org/abs/2605.21127.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. arXiv:2103.06944.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2201.11903.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html. arXiv:2407.15018.

Qinwu Xu, Zhuoheng Li, and Jessie Salas. Robust checkpoint selection for multimodal LLMs via agentic evaluation and stability-aware ranking. *arXiv preprint arXiv:2605.18852*, 2026. URL https://arxiv.org/abs/2605.18852.

Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. TinyV: Reducing false negatives in verification improves RL for LLM reasoning. *arXiv preprint arXiv:2505.14625*, 2025. URL https://arxiv.org/abs/2505.14625.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Linyao Yang, Jian-Tao Huang, Yafei Lu, Zhenhui Jessie Li, and Guirong Xue. The emperor's new reasoning: Format imitation overshadows genuine mathematical understanding in SFT. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21087–21100, Suzhou, China, November 2025b. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1067. URL https://aclanthology.org/2025.emnlp-main.1067/.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Niu. xFinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2405.11874.

Wenbo Zhang, Hengrui Cai, and Wenyu Chen. Beyond the singular: Revealing the value of multiple generations in benchmark evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10033–10043, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.488. URL https://aclanthology.org/2026.findings-acl.488/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 55006–55021, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html. arXiv:2305.11206.

APPENDIX

CONTENTS OF THE APPENDIX

Part I Extended account of the main text **20**

A Details behind the main text **20**

A.1 Program history and positioning . . . 20
A.2 Protocol, trajectories and families . . . 21
A.3 What the strict grader saw . . . 22
A.4 What a strict score is made of: bound, scorers and validation . . . 22
A.5 What the answers were: contrasts in detail . . . 24
A.6 Why the grader saw what it saw: detail . . . 26
A.7 Recommendations and limitations: extended . . . 27

Part II Setup and protocol **28**

B Bed and training specification **28**

B.1 Contract bytes . . . 28
B.2 Decoding . . . 28
B.3 Strict extractor . . . 29
B.4 Task panel and answer spaces . . . 29
B.5 Training configuration (main Qwen3 program, except the science recipe) . . . 29
B.6 Data sources and licences . . . 29
B.7 Teacher census . . . 30
B.8 Contamination audit . . . 30
B.9 Design choices and their rationale . . . 30
B.10 Statistics . . . 30

C Datasets of the box-appending retrains **32**

D Pre-registration and audit chain **32**

Part III Measurement and validation **33**

E Judge protocol and validation **33**

E.1 Prompt . . . 33
E.2 Validation statistics . . . 33
E.3 A third judge (exploratory, post hoc) . . . 34
E.4 Contested endings and the AI reading . . . 34
E.5 Hedged answers, worst case (exploratory, post hoc) . . . 35
E.6 Relation to judge meta-evaluations . . . 35

E.7 The robust-boxed rule . . . 36

F Per-cell table and strict-score decomposition **36**

F.1 Headline contrasts . . . 36

F.2 Where the strict loss lived . . . 36

F.3 Consequences for decisions . . . 37

F.4 Sensitivity to the commitment reading . . . 37

F.5 The learning-curve dip on fresh seeds (post hoc) . . . 38

G Termination outcomes and an uninformative frozen readout **38**

H Proof of the bound and additional analyses **41**

H.1 Statement and proof of the bound . . . 41

H.2 Length relative to the training distribution . . . 42

H.3 What precedes a re-emitted closer . . . 42

H.4 Prose commitments are not driven by numeric options . . . 42

H.5 Answer reliability by termination state . . . 43

H.6 A symbolic-equivalence checker can repair format, not termination . . . 43

H.7 Harness census . . . 43

H.8 A judge-free version of the protocol (exploratory, post hoc) . . . 44

H.9 Hedging . . . 45

H.10 The bound on Gemma-4-31B and a decision table . . . 45

I Re-scoring without new generation **45**

I.1 Two mechanisms behind the decisions . . . 45

I.2 Public-model rankings . . . 45

I.3 Reversals within a trajectory . . . 46

I.4 The sign rule where both counts are present . . . 47

I.5 Checkpoint selection . . . 47

I.6 Equivalence bounds for the rescue . . . 47

I.7 Answers stated before the cut . . . 48

Part IV Robustness and generality 50

J Pre-registered held-out test **50**

J.1 Outcome . . . 50

J.2 What transferred . . . 54

J.3 Second judge family (sensitivity, not registered) . . . 54

J.4 Greedy decoding (exploratory, not registered) . . . 54

J.5 Presence penalty and a doubled budget (exploratory, not registered) . . . 55

J.6 Model-card decoding at step 200 (exploratory, not registered) . . . 55

J.7 Training overlap . . . 55

J.8 Scorer invariance under answer-preserving rewrites (original program, not registered) 56

K A rerun with checkpoints every five steps **57**

L Prospective test on public models **58**

M Added students **58**

M.1 DeepSeek-R1-Distill-Qwen-7B (R1-Qwen) 58

M.2 DeepSeek-R1-Distill-Llama-8B (R1-Llama) 59

N A controlled format test on Qwen3-4B-Base **61**

O Seeds, learning rates and data in the second harness **62**

O.1 Learning rate (exploratory, post hoc) 64

O.2 Learning rate on the held-out panel (exploratory, post hoc) 64

O.3 Last checkpoints and the training-time system prompt on the held-out panel (exploratory, post hoc) 65

O.4 Training data (exploratory, post hoc) 66

P The untouched base model and a system-prompt variant **66**

Q An external science recipe with well-formatted targets **67**

R Exploratory training evaluations on Gemma-4-31B **67**

Part V Related work 69

S Positioning relative to prior work **69**

S.1 Termination in long-CoT SFT 69

S.2 Verifier error 71

S.3 Format and training state 71

S.4 State-wise evaluation 71

S.5 Statistics and validity 71

Part I Extended account of the main text

A Details behind the main text

This appendix expands §§1–6 and 8–9 of the main text, in their order, with full numbers; the cross-model evidence of §7 is detailed in Appendices J, L, M, N and R.

Terminology. The appendices use the main text's terms, with two exceptions kept from the frozen analyses: tables label delivered accuracy *committed* (comm.), since a finished response commits to an answer in its visible channel, and the gap (strict − delivered) is also called the *artifact*. Students: Qwen3-30B-A3B-Thinking-2507 (the main student), DeepSeek-R1-Distill-Qwen-7B (R1-Qwen) and DeepSeek-R1-Distill-Llama-8B (R1-Llama); Qwen3-4B-Base is a controlled format test. Registered predictions: P1–P9 (held-out items), P-A–P-C (public models), S1–S6 (R1-Qwen and R1-Llama), A1–A3 (R1-Llama addendum), Q1–Q7 (Qwen3-4B-Base); an index opens Part IV.

A.1 Program history and positioning

The program began as a study of a chat-template intervention, the *response contract*: generation begins either inside the model's thought channel (NATIVE) or after an appended empty thought closer (PRECLOSED), off-template for this thinking-only model and akin to no-thinking prompting (§2). The program was pre-registered and replicated on two fresh training seeds (§3). Under the strict grader the results were clean (§A.3): pre-closing cost a checkpoint distilled from a prose-style Gemini teacher 12–27 points, the loss was absent for a sibling recipe distilled from a procedural gpt-oss teacher, it grew from zero at step 25 to full size by step 200, and it was not mediated by termination. A pre-registered data surgery that appended a gold boxed answer to 99.8% of targets lifted the PRECLOSED AIME arm by 31 points.

Then we read the generations. Under PRECLOSED the same checkpoint ends with "Answer: C" or "Result is 167."—committed, in prose, unboxed. We re-scored every generation post hoc by termination state and by the answer it commits to in its visible channel, with a gold-blind judge validated four ways (§4). At step 200 the contract effect is -14.3 points strict but only -0.95 committed (-4.4 at worst, if every hedged answer counts as uncommitted): mostly an artifact. So is the rescue: the surgery taught the model to box, not to solve. The effect's growth is real but runs the other way: at step 25 the NATIVE arm ends inside its thought on a quarter of GPQA rows, so under the visible-channel reading PRECLOSED is *ahead* by 16.6 points, and the pre-registered interaction (-17.6 committed) is mostly that early failure disappearing. Crediting boxed guesses in unfinished thoughts, the strict grader missed that failure. For the sibling recipe it reported an early cost of pre-closing of -3.9 pp, which delivered scoring puts at -9.6 pp when truncated rows deliver nothing but at -5.1 to -1.5 pp when answers stated before the cut are credited (Appendix I.7).

Two regularities of training, both consistent with a data-side origin, explain what a strict score is made of (§6). First, the early termination collapse is not imitation of long targets: the teachers' thoughts have medians of 11–12k tokens, the base model's 7k, and the step-25 Gemini student's 30k, with a quarter of GPQA rows ending inside an unclosed thought whose last lines mostly hold the requested boxed answer. Second, *where* a student states its final answer follows its training targets: the prose teacher answers inside its thought and almost never boxes, the procedural teacher boxes in both channels, and their students diverge accordingly under PRECLOSED. A box-removal twin on Qwen3-4B-Base then isolates answer format causally (§6, Appendix N; two training seeds).

The contributions are stated in §1; the numbers behind them follow. In brief, the prospective evidence: on new items with the same checkpoints, six of nine registered predictions pass, including P3 (step-25 credit), P4 (unread step-200 Gemini PRECLOSED answers) and the recipe-gap prediction P7; P5 (gpt-oss step-200 cells within ± 3 pp and with few unparsed answers) fails, and all three failures are reported (Appendix J). R1-Qwen reproduces the collapse but not the directional predictions (Appendix M.1); R1-Llama reproduces both halves, positive step-25 Gemini gaps and a negative step-200 gap wherever b exceeds a by 5 points (Appendix M.2); removing only the boxes from the targets reproduces the late error in Qwen3-4B-Base (Appendix N); and all three public-model predictions hold across twelve models (Appendix L). Exploratory Gemma training evaluations provide further cross-model evidence (Appendix R).

Answer extraction. Extraction is a known weak link: models adapt to extractor-friendly formats (Yu et al., 2025); extraction choices change multiple-choice conclusions (Molfese et al., 2025) and, with truncation, reasoning scores (Hochlehnert et al., 2025); much reported *prompt* sensitivity (Sclar et al., 2024; Polo et al., 2024; Chatterjee et al., 2024; Liu & Chu, 2026) disappears under an LLM correctness judge (Hua et al., 2025); and five extraction rules reorder five Qwen3 and DeepSeek-R1 models, whose 3–7% of outputs with unclosed thinking cannot be read reliably (Jo et al., 2025). Trained verifiers (Chen et al., 2025a; Liu et al., 2025b) and symbolic checkers (Appendix H) target format, not termination; Long et al. (2025) separate performance under format adherence from performance regardless of it, the closest precedent to our commitment/correctness split. In reinforcement learning (RL), rule-based verifiers miss more correct answers as policies strengthen, and model-based ones can be hacked (Huang et al., 2025; Xu et al., 2025). We show that extraction error depends on SFT training state and on the format of the training targets.

Thinking, no-thinking and budgets. Hybrid Qwen3 models render their non-thinking mode as an empty thought block (Yang et al., 2025a), and a distilled reasoning model prompted to skip thinking (with a dummy thought) is competitive under matched budgets (Ma et al., 2025); our PRECLOSED arm applies this to SFT checkpoints. Budgets change rankings (Sun et al., 2025) and trade answer rate against accuracy (Jurayj et al., 2025), inference regimes hide heterogeneity (Hariri et al., 2026), and gains from more thinking can be metric artifacts (Ghosal et al., 2025); we treat hitting the budget and ending inside the thought as termination outcomes, not wrong answers. Native thinking tokens (Guo et al., 2025), descended from chain-of-thought prompting (Wei et al., 2022), define the object we study; we make no claim about the traces themselves (Jacovi et al., 2024).

Learning dynamics of long chain-of-thought (long-CoT) SFT. Across SFT runs with more epochs, termination tracks accuracy, which Kopiczko et al. (2026) attribute to unterminated outputs yielding no answer; along one run, we show, a strict grader first credits guesses in unfinished thoughts, then misses prose answers, so its error changes sign. Checkpoint studies report dips with length surges and missing thought closers (Ren et al., 2026; Luo et al., 2025) and an early out-of-distribution peak (Jin et al., 2025); recipe studies rank data by strict scores (Guha et al., 2025; Yeo et al., 2025); fine-tuning alters the link between traces and answers (Lobo et al., 2025) and may mostly teach format (Zhou et al., 2023; Yang et al., 2025b); and some "SFT fails to generalize" results are frozen-template artifacts (Lin et al., 2025). Such dips occur where a strict grader errs most, which checkpoint selection (Xu et al., 2026) must survive.

A.2 PROTOCOL, TRAJECTORIES AND FAMILIES

Contract. Both arms render the same user turn with the model-native template. NATIVE ends the prompt with the opened thought channel (`<|im_start|>assistant\n<think>\n`); PRECLOSED appends an empty closer (`</think>\n\n`), which this thinking-only template never emits (Appendix B). All else is identical. Three factors vary: the contract sets *where* the answer is written and read, the targets' answer format (boxed or prose) sets *how*, and the training step sets *whether* the response finishes rather than ending inside its thought or at the token budget.

Protocol. Our *original program* is AIME 2025 (30 items) and GPQA Diamond (198 items; Rein et al., 2024), finite answer sets; four fixed decoding seeds; T=1.0, top-p=1.0, top-k off; a 32,768 new-token budget, the training sequence length; vLLM (Kwon et al., 2023). T=1 samples the model's own distribution, as pass@k estimation and RL rollouts do. In exploratory probes with greedy or model-card decoding (with or without its presence penalty) the step-25 collapse persists, but no row ends unclosed, budget-hit rows often loop and, for the Gemini student, almost never hold a box (0–4%, against 3–71% for gpt-oss), so its step-25 credit nearly vanishes. Under the card's decoding the Gemini student still loops at step 200 (36–77% budget hits), so strict credits boxes in its unfinished NATIVE rows again and, on held-out GPQA, shows no contract effect (+0.4 pp) where delivered scoring puts PRECLOSED 15.6 pp ahead (Appendix J). The prompt asks for `\boxed{}` (AIME) or a final line `Answer: \boxed{A/B/C/D}` (GPQA). Each task's first item, seen during piloting, was excluded (1,808 confirmatory rows per checkpoint; 29 AIME items, so AIME intervals are wide). A *row* is one generation; a *cell* is one checkpoint, task and contract. *Strict accuracy*: take the text after the last `</think>` (else the whole output), find the last `\boxed{}` holding an integer (AIME) or a letter A–D (GPQA), compare with gold. Registered permutation

tests are item-clustered (B=10,000), with training seeds as fixed replicates; pre-registered families are Holm-corrected (Holm, 1979), never pooled. A registered test *rejects* when its Holm-adjusted permutation $p < .05$; elsewhere *significant* means that the unadjusted 95% confidence interval (CI) excludes zero, and every headline contrast that the main text calls significant survives a max-$|t|$ adjustment over the 14 headline contrasts (Appendix F).

Trajectories. The main program's trajectories are full-parameter SFT of Qwen3-30B-A3B-Thinking-2507 (Yang et al., 2025a) on one of two 100k-example long-CoT distillation recipes over the same 100,000 mathematics questions from NVIDIA's public Nemotron-Math-v2 corpus (Du et al., 2025): the **Gemini recipe**, whose targets we generated with `gemini-3-flash-preview` and a thought-prefill prompt (a gpt-oss-120b verifier judges 20.4% wrong) and whose thought or answer contains a `\boxed{}` on 0.08% of targets, and the **gpt-oss recipe**, the corpus's own gpt-oss-120b solutions (100% boxed). Both close the thought boundary on 100% of targets, and neither was filtered for correctness; identical optimizer, learning rate 10^{-5}, seeds, packing and schedule (Appendix B). The recipes are ecological, not a matched style intervention: they differ in teacher, length distribution, the share of over-length targets dropped (5.3% vs 16.2%) and possibly target correctness. Checkpoints come every 25 steps of 885 (Gemini) or 702 (gpt-oss); step 25 is inside the 5% warm-up, and up to step 200 (68% of the first epoch; 85% for gpt-oss) no example is seen twice; steps 25–200 are the early window over which the recipe gap that prompted the study was registered (Appendix O). An exploratory rerun of the Gemini trajectory saving every 5 steps (30 held-out AIME 2024 rows per cell) shows the collapse arriving abruptly: steps 5–15 look like the base model ($\leq 6.7\%$ budget hits, strict = committed), and at step 20 90.0% of NATIVE rows hit the budget (Appendix K).

Families (Table 6). The **seed-42** family tests the *interaction*, the PRECLOSED−NATIVE strict-accuracy difference at step 200 minus that at step 25, on the original trajectory (training seed 42); the **fresh-seed replication** tests it averaged over two fresh training seeds, a four-test Holm family. Two single-variable **box-appending retrains** append "`The final answer is $\boxed{·}$.`" to the answer segment of the Gemini targets only, leaving thinking and inputs untouched, and retrain with identical arguments: the *heuristic box* (62.19% of targets) with restatements captured from the teacher text (about 27% are long expressions), the *gold box* (99.77%) with the source gold answers. The confirmatory estimand is $\Delta = R_{\text{surgery}}(200) - R_{\text{Gemini}}(200)$ with contract reliance R = NATIVE−PRECLOSED strict accuracy, rejection direction $\Delta < 0$, Holm-corrected over the two tasks of each surgery; the gold-box protocol predates the heuristic-box results. The **second harness** scores three training seeds per recipe at eight checkpoints with a 65k-token budget and its own strict extractor (nemo-skills), native contract only. Each family was frozen, audited and independently recomputed before any number was read (Appendix D). Unless noted, contrasts are PRECLOSED−NATIVE ($= -R$), recipe gaps gpt-oss−Gemini, in percentage points (pp).

A.3 What the strict grader saw

All registered evaluations completed (Table 3). The Gemini GPQA interaction replicated on both fresh training seeds (-14.15 pp; seeds -10.66 and -17.64), as did the AIME interaction (-18.10 pp); neither gpt-oss family rejected; the seed-42 family rejected for Gemini (AIME -26.72, GPQA -12.31 pp). Termination did not explain it: at the fresh-seed step-200 cells where accuracy fell by 12–17 pp, the budget-hit rate moved by -6.0 to $+0.1$ pp (Tables 12 and 16). Of the recipe statistics we compared, only boxing differed by orders of magnitude, so we hypothesised that SFT distils where the answer is stated, and tested it: the gold box contracted reliance with rejections on both tasks and lifted the PRECLOSED AIME arm from 3.45 to 34.48% ($+31.03$ pp); the heuristic box dropped the NATIVE arm from 26.72 to 6.90%.

A.4 What a strict score is made of: bound, scorers and validation

Every response falls into one of three termination states (Figure 2): it hit the budget (B), it emitted end-of-turn while its thought channel was still open (U, possible only under NATIVE), or it closed the channel and ended (C). Write π for the event that the strict rule finds a well-formed `\boxed{}` answer in the text it reads (the answer channel if a closer exists, otherwise the whole output), y_π for that answer, and y_c for the answer a response in state C commits to in its visible channel, in

Table 3: Pre-registered strict-accuracy tests (strict: last `\boxed{}` after the closer; AIME25: AIME 2025, 29 items; GPQA: GPQA Diamond, 197 items; Gemini, gpt-oss: recipes named for their teachers; pp: percentage points). Replication: the *interaction*, PRECLOSED−NATIVE (generation after vs. inside the thought) at step 200 minus step 25, mean of two fresh seeds; Holm over four tests. Heuristic, gold box: surgeries appending a heuristic or gold `\boxed{}` answer to Gemini targets; $\Delta = R(\text{surgery}) - R(\text{Gemini})$ at step 200, contract reliance R = NATIVE−PRECLOSED; Holm over tasks. Brackets: unadjusted item-cluster 95% CI; reject: Holm-adjusted permutation $p < .05$, so a CI may exclude 0 without rejection (gpt-oss GPQA).

family	task	estimate [95% CI] (pp)	Holm p	reject
replication, Gemini	AIME25	-18.10 [$-29.31, -7.76$]	.003	**yes**
	GPQA	-14.15 [$-18.34, -9.90$]	$<.001$	**yes**
replication, gpt-oss	AIME25	$+2.59$ [$-6.47, +12.07$]	.624	no
	GPQA	$+3.55$ [$+0.06, +7.04$]	.098	no
heuristic box (62%)	AIME25	-24.14 [$-38.79, -11.21$]	$<.001$	**yes**
	GPQA	-3.55 [$-9.14, +2.03$]	.107	no
gold box (99.8%)	AIME25	-17.24 [$-25.86, -9.48$]	$<.001$	**yes**
	GPQA	-9.14 [$-14.09, -4.19$]	$<.001$	**yes**

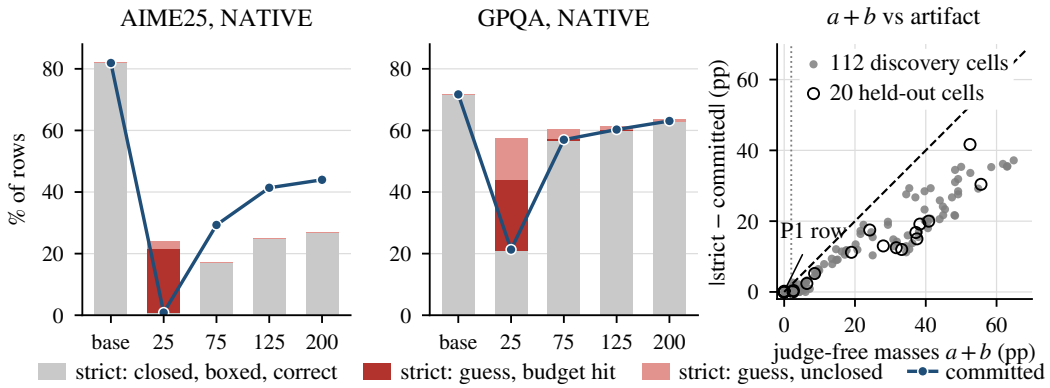

Figure 7: Left, centre: strict (last-`\boxed{}`) accuracy of the seed-42 Gemini run (NATIVE: generation starts inside the thought; base: before SFT) by training step, split by equation 1: correct boxed answers after a closed thought (grey), boxed guesses in budget-hit (dark red) or unclosed (light red) thoughts; line: delivered accuracy (judge-read visible answer). Right: |strict−committed| vs. the judge-free counts $a{+}b$ of equation 3 in 112 discovery cells (filled) and 20 held-out cells (hollow); dashed: equality; dotted: $a{+}b = 2\,\text{pp}$; P1 row: the held-out cell violating the registered judge-free bound (prediction P1, Appendix J). AIME25: AIME 2025; GPQA: GPQA Diamond; Gemini: recipe named for its teacher; pp: percentage points.

any form. With g the gold answer, strict and delivered accuracy decompose exactly as in equation 1 and equation 2 of §4 (read Delivered as Committed). The second and third terms of equation 1 credit guesses in unfinished or unclosed thoughts; the second term of equation 2 is answers stated without a box. Their difference, the *artifact* $\text{Strict} - \text{Delivered}$, is bounded on each side by a judge-free count plus the parser–judge disagreement, as in equation 3 of §4. Up to δ, strict can exceed committed only through guesses read out of unfinished states (mass a) and fall short only through unparsed answers (mass b); both come from finish reasons, the closer flag and the parser. Parser–judge disagreement is rare: δ is 0.21% here (0.30% of closed, parsed rows). In all 112 cells (28 checkpoints $\times$ 2 tasks $\times$ 2 contracts) $|\text{artifact}| \le a{+}b$ (the one-sided bound $\le a$ fails in two, by one row); the larger mass gives the sign of 91 of the 97 nonzero artifacts; the median $|\text{artifact}|$ is half of $a{+}b$; and in the 20 cells where that sum is below 2 points the artifact is at most 1.14 (the *2-point rule*; Figure 7). The pre-registered held-out panel (new items: AIME 2024 and the 250 GPQA items outside Diamond, "GPQA-main"; Appendix J) repeats all three regularities except in one cell ($a{=}b{=}0$, one disagreement row). On twelve public reasoning models decoded with their recommended settings, predictions registered before any output held in every eligible cell (size bound 42/42, sign 14/14, 2-point rule 16/16); descriptively, strict grading inverted 13 of their 66 pairs on GPQA-main (6 survive simultaneous intervals; 8 are marginally robust in the registered analyzer,

Table 4: **Only the delivered-accuracy reader ignores unfinished thoughts and reads unboxed answers.** Strict: the frozen last-`\boxed{}` rule; robust boxed: the same rule made tolerant of formatting; committed: the answer a judge that does not see the gold answer reads in the visible channel (the answer channel after the thought closer), with no delivered answer assigned to either budget-hit or unclosed responses.

scorer	what it reads	boxed guess in an unfinished thought	answer stated without a box
strict	the last `\boxed{}` after the last `</think>`, else anywhere in the output	credited if correct	never credited
robust boxed	as strict, with balanced braces and `\text{}` wrappers stripped	credited if correct	never credited
committed	the single final answer in the last 3,000 characters of the visible channel	never credited	credited if correct

7 in our recomputation; Appendix I), mostly for models that state the letter unboxed (Appendix L). Writing $\pi_s = P(\pi \mid s)$ and $\rho_s = P(y_\pi{=}g \mid \pi, s)$, the same identity factorises as

$$\text{Strict} = \textstyle\sum_{s\in\{\text{B,U,C}\}} P(s)\,\pi_s\,\rho_s, \tag{5}$$

so any difference between two strict scores is allocated exactly, by a Shapley decomposition (Shapley, 1953) over the three factors of each state, to a change in the *state mix* (termination), in *parse within state* (format), or in *correctness given parse* (reliability).

Three scorers (Table 4). **Strict** is the frozen rule; its whole-output fallback is also the default reading in the training and evaluation harnesses we inspected (Appendix H). **Robust boxed** takes the last `\boxed{}` with balanced braces and strips `\text{}` wrappers; it moves no original-program contract contrast by more than 2.6 pp. **Committed** asks an instruction-tuned judge to read the last 3,000 characters of the visible answer channel and return the single final answer the response commits to, or NONE; states B and U commit to nothing by definition. This matches a reasoning parser: the template puts the answer after the closer, and NeMo-Skills with `parse_reasoning=True` (not its default) grades an unclosed row as empty (Appendix H). The judge sees neither the gold answer, the strict parse, nor the problem. Any output other than a whole integer (AIME) or one option letter (GPQA) is never correct (11 of 39,998).

Validation. Two judges from different families, Qwen3-30B-A3B-Instruct-2507 and gemma-3-12b-it, run greedily on all 40,384 terminated generations, agree on 99.44% of rows and with the strict parse where it exists (36,884 rows) on 99.66% and 99.75%; on the 3,500 contested rows (terminated, no strict parse) they agree on 95.91% and mark 64.8% and 64.4% correct. Three checks argue against invented commitments: the judged letter or integer occurs literally in the judged tail for 99.80% of answers, and the 78 others are correct 15% of the time, below chance; contested answers are about as often correct as boxed ones in the same cells (fresh-seed Gemini PRECLOSED GPQA: 64.7 vs 64.6%); and under 26,785 answer-preserving rewrites of 6,561 verified rows the judge's verdict is unchanged on 99.97% (original program; Appendix J). If every hedged answer counts as uncommitted, no artifact changes sign or loses significance (largest shift 4.19 pp); a non-blind AI reading of 100 contested rows (no human labels) found 93 exact extractions (Appendix E). All results below use the first judge; recomputing every delivered-answer contrast with the second judge moves no estimate by more than 0.9 pp, and a third judge keeps every artifact's sign (exploratory).

A.5 What the answers were: contrasts in detail

The contract effect. Late in training the strict grader misses prose answers; early, it credits guesses in unfinished thoughts (Figure 4; its 14 max-$|t|$ contrasts: Figure 8, Table 10; all committed analyses are post hoc). On the fresh training seeds, the Gemini GPQA PRECLOSED$-$NATIVE gap at step 200 is -14.34 pp strict and -0.95 pp committed (not significant), a significant artifact of -13.39 pp. On AIME the step-200 gap is -15.09 strict and -0.86 pp committed (artifact -14.22, significant). The seed-42 family agrees (strict$\rightarrow$committed, AIME: $-23.28 \rightarrow -6.03$ pp; GPQA: $-12.82 \rightarrow -0.89$ pp; neither committed gap significant); on AIME the 29-item intervals exclude the strict estimates yet admit losses of up to 15 pp. The pre-registered *interaction* differs: it is -14.15 strict and a significant -17.58 pp committed—larger, not smaller. At step 25 the NATIVE arm ends inside

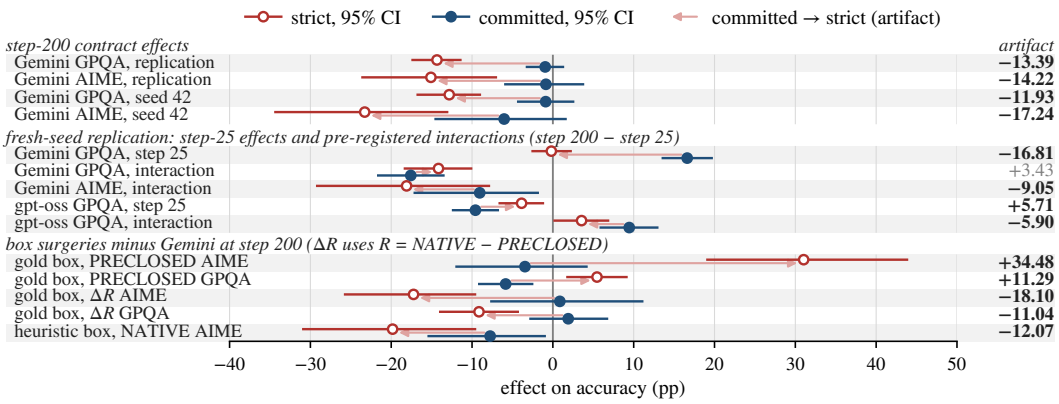

Figure 8: Figure 4 restricted to the 14 contrasts of Table 10 (the other three: Appendix F). Strict (last-`\boxed{}`; red) and committed (judge-read visible answer; blue) estimates of the headline contrasts, item-cluster 95% CIs, PRECLOSED−NATIVE (generation after minus inside the thought) unless stated. Replication: two fresh training seeds; seed 42: the original run; interaction: step 200 minus step 25; box-appending retrains: Gemini targets given a gold or heuristic box, minus the Gemini recipe; ΔR: change in contract reliance R = NATIVE−PRECLOSED (negative: smaller contract effect). Right: artifact (strict−committed), bold if its CI excludes 0. AIME: AIME 2025 (29 items); GPQA: GPQA Diamond (197); Gemini, gpt-oss: recipes named for their teachers; pp: percentage points. Post hoc; values: Table 10.

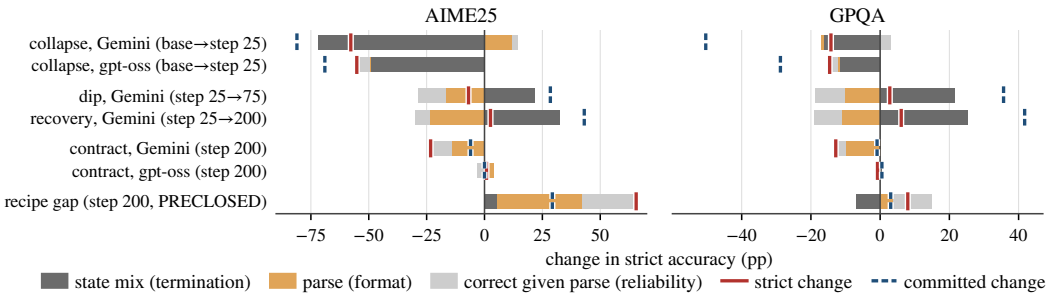

Figure 9: What moves a strict (last-`\boxed{}`) score: each strict difference (red tick) split exactly by equation 5 into state mix (termination), parse within state (format) and correctness given parse (reliability); dashed blue: the difference under delivered scoring (judge-read visible answer). Collapse: base→step 25; dip: step 25→75; recovery: step 25→200; contract: NATIVE→PRECLOSED (generation starting inside→after the thought); recipe gap: Gemini→gpt-oss (recipes named for their teachers). Seed-42 runs, NATIVE unless stated; AIME25: AIME 2025; GPQA: GPQA Diamond; pp: percentage points; all contrasts: Table 11.

its unclosed thought on 21–24% of GPQA rows (Table 12), whereas under PRECLOSED the whole output is visible, so the step-25 gap is a significant +16.62 pp in favour of pre-closing. The strict grader read the step-25 NATIVE arm's unclosed and unfinished thoughts and scored the two arms as equal (−0.19 pp): the "growth" is mostly an early closure failure disappearing, not a step-200 loss appearing. Reading unclosed rows on their whole output, as we did first, gives +3.24 pp at step 25 and an interaction of −5.33 pp, whose artifact turns to −8.82 pp; the held-out test registered the visible-channel reading before generation, and no step-200 contrast moves by more than 1.14 pp (at most 2.0% of a step-200 cell's rows end unclosed).

An early effect whose size depends on the reading. The gpt-oss recipe, which shows no step-200 artifact on this bed (but see prediction P5, Appendix J), has a real early contract effect. At step 25 its PRECLOSED arm hits the budget on 50–53% of GPQA rows against 35–39% for NATIVE (three training seeds), and its delivered accuracy is a significant 9.6 pp lower (seed 42: 9.3). The strict

Table 5: Where the final answer is stated (%; answer chars: mean characters after the last closer </think>). Teachers: 3,000 targets per recipe (Gemini, gpt-oss: named for their teachers); students (seed 42, step 200) and base model: closed GPQA rows. Phrase: "Final Answer" in the thought; box: a \boxed in the thought or answer segment; re-emits closer: a PRECLOSED rollout (sampled output) generates its own closer; NATIVE/PRECLOSED: generation starts inside / after the thought; —: not applicable; bold: largest Gemini deviations.

		NATIVE rollout or training target			PRECLOSED rollout	
recipe	source	phrase, thought	box, thought	box, answer	re-emits closer	answer chars
Gemini	teacher	58.4	**0.1**	**0.0**	—	—
	student, step 200	64.9	42.7	99.7	**14.2**	**21,747**
gpt-oss	teacher	99.5	98.1	100.0	—	—
	student, step 200	92.0	91.2	100.0	94.5	2,437
—	base model	31.9	32.2	100.0	100.0	2,647

grader reported -3.9 and -1.8 pp, having credited boxed guesses in both arms' truncated outputs; by step 200 the effect is gone (-0.13 pp), so the committed interaction is a significant $+9.45$ pp; it recurs on held-out GPQA-main (Appendix J). Its size depends on how answers stated in truncated rows are counted (Appendix I.7).

Surgeries. The gold box taught the model to box (strict parse $10 \rightarrow 52\%$ on PRECLOSED AIME) without raising delivered accuracy; the pre-registered ΔR is -17.24 and -9.14 strict, $+0.86$ and $+1.90$ pp committed. Its GPQA committed loss (-5.84 pp, PRECLOSED) vanishes if hedged answers count as uncommitted, and may come from gold boxes that contradict the target's own conclusion on about a fifth of rows (Appendix C). The heuristic-box native collapse is 39–52% real across three judges, the rest a parser rejecting boxed equations (Appendices E, C).

The dip. The seed-42 Gemini strict NATIVE AIME curve dips, $24.1 \rightarrow 17.2 \rightarrow 25.0 \rightarrow 26.7\%$; the committed one is $0.9 \rightarrow 29.3 \rightarrow 41.4 \rightarrow 44.0\%$ (Figure 7). On the two fresh training seeds, decoded under one engine configuration, strict accuracy falls by 9.1 points from step 25 to 75 while delivered accuracy rises by 31.5 (Appendix F.5). In the second harness (Appendix O), whose extractor parses at most 68% of Gemini AIME rows, the recipe gap that prompted this study (trapezoidal mean accuracy over steps 25–200, both tasks) is 28.75 pp strict, 19.55 committed.

What moves a strict score. Equation 5 splits each headline difference (Figure 9; descriptive; the convention for undefined factors moves a term by at most 8.4 pp, Table 11). The collapse is mostly state mix (termination); the strict *dip* from step 25 to 75 on AIME is the two clocks crossing, termination recovering as credit vanishes and prose arrives; the step-200 contract effect is mostly format; and the AIME recipe gap under PRECLOSED at step 200 is mostly format and reliability, of which delivered scoring keeps under half.

What is real. The termination collapse (§6); gpt-oss's early termination penalty under PRECLOSED; and, at $T{=}1$, small Gemini contract effects at step 200: -6.0 to $+2.6$ pp across three seeds on AIME, indistinguishable from zero on 29 items, and about -1 pp on GPQA, or -4.4 to -5.1 pp, unadjusted CIs excluding zero, if hedged answers count as uncommitted (Appendix E).

A.6 WHY THE GRADER SAW WHAT IT SAW: DETAIL

Termination collapse outruns the trained length range. The teachers' thoughts are long but bounded: median 12,221 tokens for the Gemini teacher and 11,448 for the gpt-oss one, on training questions (Appendix B). The base model's GPQA outputs (medians 7,352 and 7,507 tokens) never hit the 32k budget. After 25 SFT steps on either recipe the median GPQA output is 19–32k tokens, 35–53% of rows hit the budget, and on AIME 77–96% do (Appendix H). No trained target exceeds the 32k budget: the student is not reproducing the teacher's length distribution but runs past every trained length. The Gemini student also often fails to close its thought. On 21–24% of NATIVE GPQA rows at step 25 the thought runs on and then ends with the prompt's requested answer form—the last 400 characters contain a \boxed{} answer in 65% of these rows—followed directly by end-

of-turn: the final-answer act performed inside the thought, then the base model's habit of stopping after an answer. At $T{=}1$ these are drifting derivations, not verbatim loops ($\leq$3% of AIME and 2–11% of GPQA budget-hit rows at step 25 fail an 8-gram test), and recovery is fast on GPQA and partial on AIME (budget hits $95.7 \rightarrow 41.4\%$ by step 200). The strict rule parses none of the base model's 33 budget-hit rows and 29–87% of the step-25 students', so at $T{=}1$ SFT teaches truncation credit; under greedy or model-card decoding it nearly vanishes at step 25 for the Gemini student but not for the gpt-oss student, whose targets box inside the thought (§6, Table 5).

Where the answer act sits follows the targets. The teachers place the final-answer act differently (Table 5): the Gemini teacher writes "Final Answer" inside its thought on 58% of targets, in the answer segment on 6%, and boxes anywhere on 0.1%; the gpt-oss teacher boxes inside the thought on 98% and in the answer on 100%. The students follow in part. Under NATIVE both close the thought and box the answer on GPQA, like the base model and unlike the Gemini teacher (on AIME the Gemini student boxes 65%). Under PRECLOSED they diverge. The base model and the gpt-oss student treat an empty thought as an instruction to think anyway: they reason in the answer channel, re-emit the closer and box. The Gemini student re-emits the closer on 14% of GPQA rows and 51% of AIME rows, otherwise leaves its derivation in the answer channel, and on AIME ends like its teacher's text: in prose. Robustness to the strict grader is a habit—reason, close, box—not an ability; the Gemini recipe, whose teacher answers inside its thought and rarely boxes, erodes it, most on PRECLOSED AIME. Self-distillation, which removes the teacher shift, shows neither collapse nor artifact, while mixing the two teachers dilutes but does not remove it (exploratory; Appendix O). The grader problem recurs on Gemma-4-31B, but there format drift is largest for the recipe whose targets box, so target format is not its only source: read as in our bed, checkpoints distilled from a boxed-target recipe score 15–38 points lower strict than committed on GPQA in both arms, and a rule that also reads the closed thought turns a -2.9 pp committed contrast into -16.0 pp (Appendix R).

Two biases, two clocks. A strict score thus reports the training clock, and equation 1 explains the shapes of §A.3 (Table 15). The credit terms peak with the collapse, at step 25; the unparsed-answer term of equation 2 appears once the student adopts the teacher's answer form in the channel the contract exposes, between steps 25 and 75 on AIME ($36 \rightarrow 79\%$ of closed PRECLOSED rows unparsed) and on GPQA ($8 \rightarrow 25\%$, then 16% by step 200). At lower learning rates both clocks run later, and both recur on held-out items. At step 885 of the lower-rate runs (65k harness, AIME 2025) delivered accuracy exceeds strict by 23.3 ($5{\cdot}10^{-6}$) and 15.8 points ($2{\cdot}10^{-6}$), and on the held-out panel the Gemini student's last checkpoint at 10^{-5} (step 885) shows 21.7 and 29.2 points on AIME 2024 (exploratory; Appendix O, Table 45). At $T{=}1$ a strict curve is a decaying positive bias plus a growing negative one: it dips where they cross, shows a "contract effect" wherever the negative bias differs between arms, and is "rescued" by any surgery that boxes every target. So the safeguards failed: pre-registration froze the estimand, not its validity; fresh seeds reproduced a recipe property; termination diagnostics missed a loss in closed rows; a strict grader detects appended boxes.

A.7 RECOMMENDATIONS AND LIMITATIONS: EXTENDED

Report the states, not a sum. Per checkpoint and contract, report the termination state, visible-channel commitment and correctness given commitment; never credit a box in an unfinished thought.

Compute the two masses before the judge. Parsed unfinished rows and unparsed closed rows cost nothing and gave the sign of 91 of the 97 nonzero artifacts here, of 173 of 187 across the 238 cells of Figure 3 (99 of 110 where both counts are nonzero) and of 35 of 43 where both reach 2 points (Appendix I); when their sum was below 2 points the artifact never exceeded 1.14 here (the 2-point rule, §4). Above it, read the answers with a validated judge (18 GPU-minutes per 40k rows here); a judge-free version and its warning rate: Appendix H.

Treat the answer act as a data property. Targets should state the final answer explicitly and consistently in the answer segment; a surgery must check that an appended answer agrees with the reasoning; check commitment, closure and format in both contracts before reading a dip as ability.

Committed-answer scoring relies on LLM judges; the only reading of contested rows is a 100-row non-blind reading by an author-supervised AI agent (no human labels) that may share their blind spots, and a judge could be lenient on hedged prose (worst case: Appendix E). Committed analyses outside the registered held-out, public-model, R1-Qwen, R1-Llama and Qwen3-4B-Base

predictions are post hoc; the visible-channel rule was fixed after a first whole-output reading (both reported). The main Qwen3 program's confirmatory decoding samples at $T{=}1$; under greedy or model-card decoding the Gemini recipe's early credit nearly vanishes, and under the card's its termination does not recover by step 200 (Appendix J). The main Qwen3 program compares two ecological recipes and cannot isolate a causal effect of style; answer format alone is isolated causally, in Qwen3-4B-Base (Appendix N; two training seeds). R1-Qwen and R1-Llama add prospective training tests, Gemma-4-31B adds exploratory training evaluations, and twelve public models test scoring at released checkpoints; these are complementary designs, not matched replications across all backbones. The held-out test changes items, not checkpoints. The confirmatory grid is steps 25–200; the last checkpoints (885 and 702) were evaluated only exploratorily, on the held-out panel (Table 45) and, at lower rates, in the 65k harness (Appendix O). A model needing more budget is penalised by every scorer that gives no credit to unfinished rows. Judges were validated only on integer and letter answers. The tightness of equation 3 and the 2-point rule (§4) are empirical regularities of the 238 pooled cells of Figure 3, where the bound's judge-free form fails in one (the held-out cell of prediction P1), and of the 60 Gemma-4 cells.

Part II Setup and protocol

B BED AND TRAINING SPECIFICATION

Table 6: Families of the main Qwen3 program. All use Qwen3-30B-A3B-Thinking-2507 as the student, the same 1,808 confirmatory rows per checkpoint (both contracts, four decoding seeds) unless noted, and a 32,768-token budget. Pre-registration documents call these families QX, TSX, BX, BXP, P0-C and FFX (rows 1–6) and the Gemini and gpt-oss recipes g1 and o1. Gemini, gpt-oss: distillation recipes whose targets come from a Gemini-family or a gpt-oss teacher; edited Gemini: Gemini targets with a heuristic or gold `\boxed{}` answer appended (the box-appending retrains); steps: optimizer steps of the evaluated checkpoints, (8) meaning eight; native only: the NATIVE contract alone, in which generation starts inside the thought (PRECLOSED starts it after an appended empty thought closer); Holm-k: Holm correction over k tests; anchor: a reference run, not pre-registered. family: a registered set of tests; original trajectory: the first training run of each recipe (seed 42); science: a separate recipe with boxed targets (Appendix Q), trained with four model seeds and evaluated at its final checkpoint only; 65k budget: 65,536 tokens; status: pre-registered or exploratory.

family	design	recipes	training seeds	steps	status
seed 42	original trajectory	Gemini, gpt-oss	42	25, 75, 125, 200	pre-registered
replication	fresh training seeds	Gemini, gpt-oss	271828, 314159	25, 200	pre-registered (Holm-4)
heuristic box	heuristic answer suffix	edited Gemini	42	25–200	pre-registered (Holm-2)
gold box	gold answer suffix	edited Gemini	42	25–200	pre-registered (Holm-2)
second harness	65k budget, native only	Gemini, gpt-oss	42, 271828, 314159	25–200 (8)	pre-registered
science recipe	separate well-formatted recipe	science	4 model seeds	final only	exploratory
base	untouched base model	—	—	—	anchor

B.1 CONTRACT BYTES

The NATIVE prompt ends with `<|im_start|>assistant\n<think>\n`; the PRECLOSED prompt is the identical byte prefix followed only by `</think>\n\n`. Evaluation prompts contain no system turn. The model is thinking-only: its card states that it supports only thinking mode, and its template has no switch for a non-thinking mode and always opens the thought. PRECLOSED is therefore an off-template prompt for this model. It follows the no-thinking rendering of the hybrid Qwen3 models (Yang et al., 2025a), which write `<think>\n\n</think>\n\n` (one line break more than our prompt), and the "NoThinking" prompting of Ma et al. (2025); the base model ignores this request, reasons anyway and re-emits the thought closer before boxing its answer (Table 5).

B.2 DECODING

vLLM, tensor-parallel 8, temperature 1.0, top-p 1.0, top-k disabled, maximum 32,768 generated tokens, special tokens retained in the returned text, four decoding seeds (9101, 9201, 9301, 9401),

both tasks and both contracts inside the same loaded engine. Each checkpoint yields 1,824 rows in 92 atomic paired shards; 1,808 rows are confirmatory.

B.3 STRICT EXTRACTOR

Let T be the returned text. If `</think>` occurs in T, the answer segment is the text after its last occurrence; otherwise it is T. A trailing end-of-turn string is removed. The last `\boxed{...}` with no nested braces whose content, after stripping `$` and commas, is an integer (AIME) or a single letter A–D (GPQA) is the parsed answer. Rows without a parse count as incorrect.

B.4 TASK PANEL AND ANSWER SPACES

The two tasks were inherited from the evaluations that preceded this program: the response contract was first noticed in archived AIME-2025 and GPQA-Diamond evaluations, the exploratory Gemma-4-31B runs that followed (Appendix R) used the same items, and every pre-registered family (Table 6) kept them. MATH (Hendrycks et al., 2021) was never part of this panel, and we report no MATH result; its absence limits scope and is not a finding. The held-out test (Appendix J) froze the strict rule and the judge protocol before generation; both are defined only for integer and letter answers, and its new items (AIME-2024, GPQA-main) have the same answer types. Both answer spaces are finite (an integer from 0 to 999, or one of four option letters), and each scorer is built for them: the strict rule parses only an integer or a single letter; the judge is told the answer type, returns at most 8 tokens and is accepted only as a whole integer or a single letter (Appendix E); and committed correctness is an exact match with the normalised gold answer. On free-form answers such as MATH's, strict and delivered scoring would each need an equivalence step (Appendix H applies one to the strict side), and the judge would need its own prompt and validation. Equations 1 and 5 hold for any extraction rule and any answer-equality test, since they only split strict-correct rows by termination state and parse outcome, and equation 3 holds with δ replaced by the share of rows that are closed and parsed and on which strict and delivered scoring disagree about correctness; the tightness of the bound and every evaluation result in this paper are measured on these two answer spaces only.

B.5 TRAINING CONFIGURATION (MAIN QWEN3 PROGRAM, EXCEPT THE SCIENCE RECIPE)

Full-parameter SFT of Qwen3-30B-A3B-Thinking-2507 with the model-native thinking template and the training-time system prompt "You are a helpful assistant."; Adam ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$), peak learning rate 10^{-5} with cosine decay to 10^{-6} and 5% warm-up, weight decay 0.1, gradient clipping 1.0, bf16, per-token loss, MoE auxiliary loss coefficient 0.001; global batch 128 packed sequences of 32,768 tokens (micro-batch 1, padding-free packing); rows longer than 32,768 tokens under the training template dropped (5,334 Gemini and 16,162 gpt-oss targets, 5.33 and 16.16%, by a full census); 1% of rows (1,000), taken by row position, held out as a validation split and not trained on, so the Gemini student trained on 93,714 rows and the gpt-oss student on 82,994, and because the two files order the same 100,000 questions differently their validation splits share 7 questions; targets were not filtered for correctness (the source corpus marks 45,832 of the gpt-oss targets as verified); three epochs; data seed and model seed 42 (271828 and 314159 for the fresh-seed trajectories); checkpoints every 25 steps; tensor/pipeline/expert parallelism $4/2/4$ on 32 GPUs. The gpt-oss and surgical trajectories differ from the Gemini trajectory only in the dataset and output paths. The scheduler reported all four fresh-seed training jobs as failed, because of an error in distributed-process teardown at the end of the run; each had already saved every checkpoint from step 25 to step 200 and trained to at least step 700 (the Gemini jobs to step 885). An amendment written after training and before any benchmark output of these seeds existed (Appendix D) accepts only their steps 25–200, the pre-specified grid, and makes no end-of-training claim for them; the seed-42 trajectories, which completed, are reported separately throughout and agree with them (§A.3, §5).

B.6 DATA SOURCES AND LICENCES

Questions and gpt-oss targets: `nvidia/Nemotron-Math-v2` (Du et al., 2025), its high-reasoning files, in which the gpt-oss-120b solutions were generated by the corpus authors; its AoPS portion is released under CC BY 4.0 and its StackOverflow and MathGenSelect portions

under CC BY-SA 4.0. The Gemini targets were generated by us for the same questions with `gemini-3-flash-preview` through a commercial API gateway, with the assistant turn prefilled by the opening of a `<thought>` block that tells the model to avoid repeating itself and to end its thinking with `</thought>` before answering. Their "thought" is therefore a prompted pseudo-thought in the response channel, not native thinking; our renderer maps it to the student's `<think>...</think>`. At most two generations were drawn per question, and the first was used for 99,957 of the 100,000 rows. A gpt-oss-120b verifier marks the used generation correct on 78,553 rows (78.6%) and incorrect on 20,385 (20.4%); the recipe was not filtered on it, so its answers can be wrong. Questions are matched across the two files by their normalised text. Training used ms-swift 3.12.0.dev0 (commit `d46a8e42`, with local modifications) and Megatron-Core 0.15.0. Science recipe: 16,526 rows (about 24M tokens) selected by our filtering pipeline from the science split (`SFT/science`) of `nvidia/Llama-Nemotron-Post-Training-Dataset` v1.1 (Bercovich et al., 2025; CC BY 4.0; responses by DeepSeek-R1 and Qwen-2.5-32B-Instruct as released); no response was regenerated. No code data is used. Evaluation: AIME 2024 and 2025 and GPQA (Rein et al., 2024).

B.7 Teacher census

A sample of every 33rd target (3,000 per recipe with a closer; all targets, including the over-length ones that training dropped, 5.3% of Gemini and 16.2% of gpt-oss targets), tokenized with the student's tokenizer: Gemini thoughts have a median of 12,221 tokens (90th percentile 27,137) and answer segments 877; gpt-oss thoughts 11,448 (42,709) and answers 968. A "Final Answer" phrase occurs inside the thought on 58.4% (Gemini) and 99.5% (gpt-oss) of targets and in the answer segment on 6.1% and 3.6%; `\boxed{}` occurs in the thought on 0.1% and 98.1%, and in the answer on 0.0% and 100.0%. The seed-42 students' protocols at step 200 (closed rows): under NATIVE the answer is boxed on 99.7% (Gemini) and 100% (gpt-oss) of GPQA rows and 64.7% and 100% of AIME rows; under PRECLOSED the closer is re-emitted on 14.2% / 94.5% of GPQA rows and 51.5% / 98.8% of AIME rows, and the mean length of the answer segment (the text after the last closer, which the strict rule reads) is 21,747 / 2,437 characters on GPQA against 2,647 for the base model, which re-emits the closer on 100% of GPQA rows. Under PRECLOSED that segment contains a box on 84.0% / 99.9% of GPQA rows and 19.1% / 100.0% of AIME rows. How often the Gemini student's closed rows lack a strict parse depends on the task as well as on the contract: across its three training seeds at step 200 it is 74–83% (PRECLOSED AIME), 37–51% (NATIVE AIME), 16–20% (PRECLOSED GPQA) and below 1% (NATIVE GPQA). The strict rule parses none of the base model's 33 budget-hit rows (all AIME: 15 NATIVE, 18 PRECLOSED; held-out panel: none of 16).

B.8 Contamination audit

A streaming lexical detector (Unicode NFKC, case folding, LaTeX-delimiter removal; flags an exact normalised match, a substring in either direction, or at least three shared 8-gram shingles with benchmark containment ≥ 0.25 and sequence ratio ≥ 0.55; thresholds frozen before the first scan) was run over both 100,000-question training corpora against the 30 AIME-2025 and 198 GPQA-Diamond questions: no exact match; two near-match flags on one training question present in both corpora, adjudicated as a different problem sharing AIME answer-form boilerplate. The same detector is applied to the held-out panel in Appendix J.

B.9 Design choices and their rationale

Table 7 gives, for each setting of the program, why it was chosen, or that it is a convention, a compute limit or inherited from the evaluations that raised the question, and where its sensitivity is shown.

B.10 Statistics

All contrasts are computed per item as the mean over seeds, then item-clustered: bootstrap CIs and permutation p values with $B = 10{,}000$. Committed-answer contrasts on the original program use the same item-cluster bootstrap (seed 1702) and are labelled post hoc. The held-out and public-model tests instead registered their delivered-answer predictions before generating any evaluation outputs (Appendices J and L), and retain their originally specified thresholds.

Table 7: Design choices and their rationale. Inherited: fixed by the evaluations that preceded the program; convention, compute: no claim of optimality. Gemini, gpt-oss: distillation recipes named for their teachers; NATIVE/PRECLOSED: generation starts inside / after the thought; strict: the last-`\boxed{}` rule; committed: the judge-read visible answer; pp: percentage points.

choice	rationale; where its sensitivity is shown
Model and data	
student model	Qwen3-30B-A3B-Thinking-2507 is thinking-only: its template always opens the thought, so PRECLOSED is a clean off-template arm; full-parameter SFT of this mixture-of-experts model fits our 32 GPUs. This is the main backbone; prospective R1-Qwen and R1-Llama training, a controlled format test on Qwen3-4B-Base and exploratory Gemma-4-31B evaluations extend model coverage (§7, Appendices M, N and R).
recipes	For the same questions, the corpus's gpt-oss-120b solutions and our Gemini-teacher targets differ by orders of magnitude in how often they box (§3): a natural, not a controlled, contrast in answer format; the box-appending retrains vary format alone.
kept as built	Neither recipe was filtered for correctness, so each stays ecological; filtering would be a further intervention (convention; Gemini verifier rates in this appendix).
over-length targets	Dropped, not truncated: truncation would train on cut-off thoughts, the failure studied here, so every trained target closes its thought; the dropped shares differ by recipe (this appendix).
system prompt	Training used a generic system string (convention); evaluation uses the template default, no system turn. Adding the string at evaluation leaves the format habit intact (Appendix P) and, on the held-out panel, the step-25 credit in place (Table 45).
Training and checkpoints	
optimizer, schedule	Convention for full-parameter long-CoT SFT. At lower peak rates both clocks recur, later (Table 43).
batch, epochs, seeds	Compute and convention (this appendix); arbitrary seeds fixed before training; a 1% split monitors loss.
checkpoint interval	Every 25 steps (convention; storage and evaluation cost); step 25, the first checkpoint, falls inside the learning-rate warm-up (Appendix B.5).
warm-up	The collapse is not tied to warm-up alone: at the lower peak rates it bottoms later, after warm-up has ended (Table 43).
confirmatory grid	Steps 25–200, the window over which the recipe gap that prompted the study was registered (Appendices A.2 and O); seed 42: evenly spaced steps 25, 75, 125, 200; fresh seeds: the two endpoints of the interaction they test.
fresh-seed runs	All four saved steps 25–200 and then failed at end-of-run teardown; an outcome-blind amendment accepted steps 25–200 only (this appendix). The complete seed-42 trajectory agrees (§A.3).
later training	Not confirmatory. At lower peak rates, and at 10^{-5} on held-out AIME 2024, delivered accuracy still exceeds strict at the last checkpoint (§7, Tables 43 and 45); late Gemma checkpoints show the same gap on GPQA (Table 47).
Evaluation	
tasks	Inherited (task panel above): finite answer spaces that the strict rule and the judge were built for, also the panel of the closest prior work (Kopiczko et al., 2026). Each first item was exposed in an earlier pilot.
prompt	Requests the form the strict rule reads (`\boxed{}`; "Answer: `\boxed{}`"), as in Kopiczko et al. (2026).
decoding	$T{=}1$, top-p 1 and four seeds per item are inherited; $T{=}1$ is the model's own sampling distribution, used for pass@k and RL rollouts. Greedy and model-card decoding keep the step-25 collapse but remove most of the Gemini student's step-25 truncation credit; under the card's decoding it still loops at step 200, where its credit returns (Appendix J).
budget	The training sequence length, so no trained target exceeds it; doubling it leaves most step-25 AIME rows at the budget (Appendix J).
second harness	Designed to compare recipes, not contracts (hence NATIVE only), in a pre-existing NeMo-Skills harness with a longer budget; it changes budget and extractor at once (Appendix O).
held-out panel	All AIME 2024 items and all GPQA-main items outside Diamond: new items, same answer types; two GPQA-main decoding seeds keep cells affordable (compute; Appendix J).
Scoring	
strict rule	Frozen in every protocol before its responses existed; its whole-output fallback is the harness default (Appendix H). The robust-boxed rule and math-verify leave the state terms (§4, Appendix H).
visible channel	The template puts the answer after the closer, where a reasoning parser such as NeMo-Skills' (`parse_reasoning=True`) reads it; registered for the held-out test; the whole-output reading is reported beside it (§5, Appendix F).
judge models	Open instruction-tuned extractors, cheap to run (§8); a second family and a reasoning judge move no conclusion (Appendix E).
judge inputs	Blind to question, gold and parse, so it extracts rather than solves; the last 3,000 characters, with shorter and longer windows tested (Appendix E); greedy; the last answer, mirroring the strict rule, with hedging reported and a worst case given (Appendix E).
Shapley allocation	Order-free and exactly additive (Shapley, 1953); the convention for undefined factors is bounded (Table 11).
registered bound (P1)	The judge-free form, the only one computable without a judge; the form with δ holds by construction (Appendix H).
2-point rule	A round threshold fixed on the original program and registered in P1; Table 18 gives other thresholds (convention).
Statistics	
item clustering	Items are the sampling unit and decoding seeds repeated measurements; resampling avoids normal approximations on 29 items (Miller, 2024; Bowyer et al., 2025).
training seeds	Two fresh seeds cannot estimate between-seed variance, so intervals cover items for these trajectories and every seed is reported; the seed-42 family is registered separately, not averaged in.
Holm, never pooled	Holm controls family-wise error without independence assumptions (Holm, 1979); each family was frozen separately, so a pooled family would be post hoc; a max-$\lvert t\rvert$ adjustment over the headline contrasts is reported (Appendix F).
interaction	Tests the seed-42 finding that the contract effect grows with training; a step-200 contrast alone would not.
heuristics	The 8-gram loop test, answer-signature windows and census samples are descriptive; only the hedging count enters a registered prediction (P9). Contamination thresholds were frozen before the first scan.

C DATASETS OF THE BOX-APPENDING RETRAINS

The gold-box-appending retrain raised the PRECLOSED arm's strict parse rate from 82.9 to 97.3% on GPQA and from 10.3 to 51.7% on AIME while commitment stayed at 98–99% and 53–58%: it taught the model to box. Its committed effect is -3.45 pp on PRECLOSED AIME (CI $[-12.07, +4.31]$), and on GPQA it *lowers* delivered accuracy in both arms (-5.84 and -3.93 pp), although the PRECLOSED loss vanishes (-1.65 pp, CI $[-5.20, +2.03]$) if every hedged answer counts as uncommitted (Appendix E). A candidate cause of the loss: the Gemini targets were not filtered for correctness (20.4% judged wrong; Appendix B), so on about a fifth of rows the appended gold box contradicts the target's own conclusion. The pre-registered ΔR on AIME and GPQA ($R =$ NATIVE$-$PRECLOSED) is -17.24 and -9.14 pp strict, $+0.86$ and $+1.90$ pp committed. The heuristic-box native strict collapse of -19.83 pp is -7.76 pp committed (CI $[-15.52, -0.86]$): the noisy suffix did some real damage on AIME, but 48–61% of the drop (across three judges) is that the model learned to box whole equations ("`\boxed{m+n = 73+4 = 77}`"), which the integer parser rejects; its strict parse rate fell from 37.1 to 12.9% while its commitment rate moved from 58.6 to 53.4%.

The heuristic-box census over all 100,000 Gemini targets found 62,188 modified rows (62.19%): 62,187 pure suffix appends and one whitespace-normalised append; zero edits to inputs or thinking. Boxed contents of at most 40 characters make up 73.45% of appended restatements (median 21, p90 62 characters); the rest are long expressions captured from the teacher text. The gold-box build joined all 100,000 targets to their source gold answers with a byte-identical question check (zero mismatches) and appended "`\n\nThe final answer is $\boxed{<gold>}$.`" to 99,771 rows (99.77%); 229 rows with an empty gold were left unchanged. Both surgical trajectories share the Gemini tokenizer, chat template and configuration files byte-for-byte and differ from the Gemini trajectory only in the dataset and output paths.

D PRE-REGISTRATION AND AUDIT CHAIN

Each confirmatory family of the program (Table 6) followed the same sequence: (i) a protocol lock naming estimands, families, rejection directions, interpretation grids and release rules, frozen before any response existed and hash-recorded; (ii) an independent design audit checking checkpoint identities, frozen inputs and the non-existence of output directories; (iii) execution with atomic shard publication and an outcome-blind completion gate that read only directory structure and manifests; (iv) a frozen analyzer whose hash was in the lock; (v) an independently implemented exact and bootstrap recomputation on a separate code path; (vi) release, followed for the fresh-seed replication by a streaming row-identity and hash audit. The locks fix two subtraction directions. The seed-42 and replication families register PRECLOSED$-$NATIVE contrasts and their step-200 minus step-25 interaction (the replication gate is signed in that direction); the box-appending retrains register the change in contract reliance, ΔR, with $R =$ NATIVE$-$PRECLOSED and rejection direction $\Delta R < 0$. Each family is reported in its registered direction, so a negative seed-42 or replication interaction means that R grew from step 25 to step 200, whereas a negative ΔR means that the surgical model ends with a lower R than the Gemini model at step 200. The gold-box freeze receipt records the non-existence of the heuristic-box gate result and analysis file and of the gold-box checkpoints at freeze time. Two parser-only errata were recorded before the delivered-answer analysis: a line-splitting bug in a text reader and the uninformative closure readout of Appendix G; neither changed a scientific number. The original program's delivered-answer re-scoring was not pre-registered; it was triggered by reading the generations after the strict results had been released and is labelled post hoc throughout. The later held-out, public-model, R1-Qwen, R1-Llama and Qwen3-4B-Base tests distinguish their registered predictions from supplementary post-hoc analyses. The held-out test has its own chain, described in Appendix J: it froze the delivered-answer rules and registered predictions on them (all but P2 and P8 involve delivered answers), recording its protocol, predictions and frozen-panel hash before generation; its analyzer was code-locked while generation was still underway, before any output was read, rather than hashed into the protocol lock as in step (iv); its one amendment is recorded with its timing. The public-model test (Appendix L) has its own chain: model list, panels, sampling and predictions P-A to P-C were hashed before any of its output existed (sha256 `35477683...`), the analyzer was code-locked before any output existed and was not changed afterwards, the two optional models entered by a completion rule fixed in advance,

and its ranking comparison was registered as descriptive. The added students have their own chains (Appendices M and N). R1-Qwen's predictions S1–S6 were frozen before any of its checkpoints existed (sha256 `d6925a56...`; its first training jobs failed at start-up without producing a step and were resubmitted unchanged); R1-Llama's verbatim copy of S1–S6 was frozen before either training job was submitted (`08c3c1c3...`), and its addendum A1–A3 after the judge had processed its base and step-25 runs but before any judge output was opened (`ee2c7805...`); Qwen3-4B-Base's Q1–Q4 were frozen before any training job was submitted (`fe90c6eb...`), and Q5–Q7, for a second training seed, before that seed's training started (`e6a168cc...`). The fresh-seed dip prediction was frozen after the seed-42 dip was known and before any fresh-seed output at steps 50–100 existed (`8cceb25a...`), and its primary analysis was designated before the step-25 re-run (`99f3845d...`; Appendix F.5). Exploratory and not registered, as labelled where they appear: the every-5-steps re-run, the learning-rate checkpoints on the held-out panel, the last checkpoints and the system-prompt variant on the held-out panel, the greedy, model-card, presence-penalty and 64k-budget probes, the third judge, the hedged worst case, the max-$|t|$ adjustment and the judge-free reader. The fresh-seed trajectories carry one deviation. Their release rule required a scheduler success status, a clean fatal-log scan and a terminal checkpoint; all four training jobs failed at end-of-run teardown (Appendix B). A salvage amendment, frozen after the training status was known and before any benchmark output of these seeds existed, replaced that rule for the pre-specified steps 25–200 only. It accepts a seed only if its configuration and data hashes match the original lock, all eight checkpoints from step 25 to step 200 are complete, training progressed to at least step 700 and the failure signatures are recorded; it prohibits end-of-training claims and the selection of checkpoints or seeds, and it requires the comparison with the completed seed-42 trajectories that §A.3 reports.

Part III Measurement and validation

E Judge protocol and validation

E.1 Prompt

System: "You are a careful grader's assistant. You extract the final answer that a response commits to. You never solve the problem yourself." User: the task type ("a competition math problem whose answer is an integer from 0 to 999" or "a four-option multiple-choice question with options A, B, C, D"), the rule to output only the answer, to output the *last* delivered answer if several conflict, and to output NONE if the response never states a final answer; then the last 3,000 characters of the visible answer channel. Greedy decoding, at most 8 new tokens. After stripping wrappers (`$`, `\boxed{}`, markdown, "Answer:"), the output is accepted only if it is a whole integer (AIME) or a single option letter (GPQA); the 11 remaining non-NONE outputs ("706/3", "140.5", "E") are out-of-domain and never correct; the original-program tables count them as uncommitted (at most one row per cell), and the held-out analyzer of Appendix J counts such outputs as commitments. The judge's letter or integer occurs literally in the judged tail for 39,797 of 39,875 answers (99.80%); the 78 exceptions, concentrated in responses that conclude with a numeric value on a multiple-choice item, are correct 12 times (15%), and scoring them as NONE changes no contrast in Table 10 by more than $0.13\,\mathrm{pp}$ (one GPQA row of 788). The prompt names a finite answer space, and the judge was validated only on integers and option letters; free-form answers, whose equivalent forms differ as strings, would need their own prompt, an equivalence step after extraction and a fresh validation, which we have not done.

E.2 Validation statistics

Over the 40,384 terminated rows of all 28 checkpoints (including the excluded first items): judge agreement Qwen3-30B-A3B-Instruct-2507 versus gemma-3-12b-it 99.44%; agreement with the strict grader on the 36,884 rows where it parses, 99.66% and 99.75%; on the 3,500 contested rows (terminated, no strict parse; 832 AIME, 2,668 GPQA), judge agreement 95.91%, NONE rates 2.8% and 4.3%, and correct-answer rates 64.8% and 64.4%. The first judge shares the students' model family (Qwen3), which could favour their phrasing; the second, from another family, agrees with it on 99.44% of terminated rows, and recomputing every delivered-answer contrast with it moves no estimate by more than $0.9\,\mathrm{pp}$ (§4). These statistics use a first-pass normaliser that, unlike the accept

rule above, takes the first integer in an AIME output and maps a multiple-choice output without an option letter to NONE: the second judge's 4.3% is 2.4% literal NONE plus 1.9% out-of-domain outputs, mostly numbers on multiple-choice items, and the first judge's 2.8% is all literal NONE; under the accept rule the two judges agree on 99.46% of terminated and 96.11% of contested rows, and the correct-answer rates are unchanged.

E.3 A THIRD JUDGE (EXPLORATORY, POST HOC)

gpt-oss-120b, a reasoning model from a third family (that of the gpt-oss teacher), re-judged all 40,384 terminated rows with the same prompt, tail and blindness, greedily at low reasoning effort, its answer read from its final channel. On the 39,998 terminated confirmatory rows its correctness verdicts agree with the first and second judges' on 99.60 and 99.63% (first with second: 99.76%; Table 8); it disagrees most on contested rows, where it returns NONE on 272 of 3,208 (the first two: 90 and 74). Over the 36 contrasts of the original program (for seed 42 and for the fresh-seed replication, each recipe's step-25, step-200 and interaction contrasts on both tasks; for each box-appending retrain, its NATIVE, PRECLOSED and ΔR contrasts on both tasks), all artifacts keep their sign (one is zero under both) and no committed estimate moves from the first judge's by more than $2.6\,\mathrm{pp}$ (second judge: 0.9). The largest move, three rows of 116, is heuristic box $-$ Gemini on NATIVE AIME at step 200 ($-7.76 \rightarrow -10.34\,\mathrm{pp}$): each response reaches the gold answer, then boxes an intermediate equation ("Final Answer: 907", then `\boxed{N = 126 + ... = 2907}`), whose value this judge returns. Of the heuristic-box native strict collapse ($-19.83\,\mathrm{pp}$, 23 rows), 9 rows are real under the first judge, 10 under the second and 12 under this one (39, 43 and 52%); under this judge the artifact is still $-9.48\,\mathrm{pp}$ (CI $[-17.24, -2.59]$).

Table 8: Agreement between the delivered-answer judges on the terminated confirmatory rows of the original program (%). *Correctness*: the same verdict under the visible-channel reading (only closed rows can be correct). *Answer*: the same extracted answer, where NONE matches only NONE and an out-of-domain output only another (both-NONE rows among the terminated: 106, 88 and 74). Q: Qwen3-30B-A3B-Instruct-2507 (first judge); G: gemma-3-12b-it (second judge); O: gpt-oss-120b (third judge). Exploratory, post hoc. Committed answer: the final answer the visible answer channel states, as the judge extracts it; strict parse: an answer the last-`\boxed{}` rule reads; original program: AIME 2025 and GPQA Diamond; NONE: no delivered answer. Terminated: ended with end-of-turn (closed or unclosed); n: rows; contested here: closed rows without a strict parse (the 3,500 of Appendix A.4 also count excluded items and unclosed rows); out-of-domain output: neither a whole integer nor a single option letter.

		correctness			answer		
rows	n	Q–O	G–O	Q–G	Q–O	G–O	Q–G
terminated	39,998	99.60	99.63	99.76	98.99	98.97	99.38
strict parse present	36,570	99.79	99.83	99.86	99.59	99.65	99.78
contested (closed, no strict parse)	3,208	97.41	97.32	98.53	93.11	92.05	95.36

E.4 CONTESTED ENDINGS AND THE AI READING

Contested rows end, for example, with "Answer: C", "(C) is the answer", "Final selection: (B)", "Result is 167.", "The sum of all such positive integers n is 49.", or a bare final line "610"; the heuristic-box model writes "The final answer is `\boxed{m+n = 73+4 = 77}`". One hundred random contested rows (40 on the first day, 60 on a second pass) were read by an AI agent under author supervision; no human labels were produced. The reading was not blind: the stored sample files it worked from show each row's gold answer and both judges' answers beside the last 320 characters of its answer channel (the judges read up to 3,000). In 93 rows the judges extracted the stated final answer; the seven remaining rows committed to a numeric value or a ranking on a multiple-choice item ("Answer: `\boxed{1}`", "5 > 2 > 1 > 3") or to a non-integer on AIME ("`\boxed{\dfrac{487}{3}}`"), where the delivered answer is genuinely ambiguous; the judges disagreed on one of them (one returned NONE). This reading is not an independent validation: its reader is itself a language model that saw the judges' answers, so it can share their conventions and blind spots (for example, whether a numeric conclusion on a multiple-choice item, or the last of several candidates in hedged prose, counts as a commitment) and may have been

anchored by the answers shown; its per-row labels were not stored with the sample files. It argues against gross extraction errors on clearly stated answers, not against leniency on hedged prose, and the delivered-answer metric has no human-only ground truth. The other checks rest on references that involve no language model (the strict parser on parsed rows, literal occurrence in the judged tail, the gold answer), but none of them separates a delivered answer from the last of several candidates, while on 32–40% of the Gemini student's contested GPQA rows that the judge answers, the last 3,000 characters name two or more different option letters in answer-like phrases (Appendix H); the rewrites of Appendix J are deterministic templates, but their rows were selected by the judges' own reading, so they test invariance, not correctness.

E.5 Hedged answers, worst case (exploratory, post hoc)

If every hedged contested row (545 GPQA and 8 AIME rows whose last 3,000 characters name two or more candidate answers in answer-like phrases) is scored as uncommitted, no headline artifact changes sign or loses significance (fresh-seed Gemini GPQA step 200: $-13.39 \to -9.96$ pp, CI $[-11.87, -8.19]$; no artifact moves more than 4.19 pp, and the Gemini GPQA interaction artifact becomes significant, $+5.90$ pp), but the step-200 GPQA committed gaps become small losses whose unadjusted CIs exclude zero (-4.38 fresh seeds, -5.08 pp seed 42) and the gold box's GPQA committed loss disappears ($-5.84 \to -1.65$ pp, CI $[-5.20, +2.03]$; Table 9). Under a harsher AIME rule that also counts boxed and bare-line integers (59 AIME rows), the already marginal AIME interaction artifact loses significance (-7.33 pp, CI $[-16.81, +2.16]$). The judges agree more often on hedged rows than on other unparsed committed rows (GPQA, first–second judge 99.08% against 95.97%, first–third 95.60% against 92.36%), but all three take the last commitment, so this agreement is not evidence against leniency.

Table 9: Hedged answers, worst case (exploratory, post hoc): committed estimate and artifact (strict − committed) of the 14 headline contrasts of Table 10 (pp). As reported: the visible-channel reading. Hedged→none: every hedged contested row (closed, no strict parse, and the last 3,000 characters name two or more candidate answers in answer-like phrases) scored uncommitted. One-sided: a harsher AIME rule that also counts boxed and bare-line integers, applied only in the cells where it moves the artifact toward zero. Bold: the item-cluster 95% CI excludes 0. Contrasts are PRECLOSED−NATIVE unless stated (generation starting after / inside the thought); replication: mean of two fresh training seeds, step 200; seed 42: the original trajectory, step 200; interaction: step 200 minus step 25; gold box, heuristic box: surgeries minus the Gemini recipe at step 200; ΔR: change in contract reliance $R =$ NATIVE−PRECLOSED; Gemini, gpt-oss: distillation recipes named for their teachers; pp: percentage points.

	committed		artifact (strict − committed)		
contrast	as reported	hedged→none	as reported	hedged→none	one-sided
Gemini GPQA, replication	-0.95	$\mathbf{-4.38}$	$\mathbf{-13.39}$	$\mathbf{-9.96}$	$\mathbf{-9.96}$
Gemini AIME, replication	-0.86	-0.86	$\mathbf{-14.22}$	$\mathbf{-14.22}$	$\mathbf{-12.50}$
Gemini GPQA, seed 42	-0.89	$\mathbf{-5.08}$	$\mathbf{-11.93}$	$\mathbf{-7.74}$	$\mathbf{-7.61}$
Gemini AIME, seed 42	-6.03	-6.90	$\mathbf{-17.24}$	$\mathbf{-16.38}$	$\mathbf{-13.79}$
Gemini GPQA, step 25	$\mathbf{+16.62}$	$\mathbf{+15.67}$	$\mathbf{-16.81}$	$\mathbf{-15.86}$	$\mathbf{-15.80}$
Gemini GPQA, interaction	$\mathbf{-17.58}$	$\mathbf{-20.05}$	$+3.43$	$\mathbf{+5.90}$	$+2.41$
Gemini AIME, interaction	$\mathbf{-9.05}$	$\mathbf{-9.05}$	$\mathbf{-9.05}$	$\mathbf{-9.05}$	-7.33
gpt-oss GPQA, step 25	$\mathbf{-9.58}$	$\mathbf{-9.90}$	$\mathbf{+5.71}$	$\mathbf{+6.03}$	$\mathbf{+5.65}$
gpt-oss GPQA, interaction	$\mathbf{+9.45}$	$\mathbf{+9.84}$	$\mathbf{-5.90}$	$\mathbf{-6.28}$	$\mathbf{-5.58}$
gold box, PRECLOSED AIME	-3.45	-2.59	$\mathbf{+34.48}$	$\mathbf{+33.62}$	$\mathbf{+31.03}$
gold box, PRECLOSED GPQA	$\mathbf{-5.84}$	-1.65	$\mathbf{+11.29}$	$\mathbf{+7.11}$	$\mathbf{+6.98}$
gold box, ΔR AIME	$+0.86$	0.00	$\mathbf{-18.10}$	$\mathbf{-17.24}$	$\mathbf{-13.79}$
gold box, ΔR GPQA	$+1.90$	-2.28	$\mathbf{-11.04}$	$\mathbf{-6.85}$	$\mathbf{-6.60}$
heuristic box, NATIVE AIME	$\mathbf{-7.76}$	$\mathbf{-8.62}$	$\mathbf{-12.07}$	$\mathbf{-11.21}$	$\mathbf{-9.48}$

E.6 Relation to judge meta-evaluations

Judge meta-evaluations such as RewardBench (Lambert et al., 2025) ask whether a judge prefers the better of two responses to a shown prompt; format and verbosity bias there means a preference swayed by presentation or length. This instrument makes no preference or correctness judgement: it is given the task type, the extraction rules and the last 3,000 characters of one response's answer channel, never the question or the reference (unlike reference-guided LLM judges (Zheng et al.,

2023)), and returns in at most 8 tokens the answer that text states; correctness is scored afterwards by exact match. The text may restate the question, but the judge's answers that do not occur literally in it are correct only 15% of the time, below chance (above). What can depend on presentation is extraction error. Format is varied by the answer-preserving rewrites of Appendix J, which change the length of the judged text by at most 81 characters and so do not test verbosity. Length is varied in a separate panel on the same 6,561 verified rows (exploratory, post hoc; same prompt and decoding): with the judged window cut to 750 or 1,500 characters or widened to 6,000 (which changes the text of 2,048 rows), the judge returns the verified answer on 99.62, 99.85 and 100% of rows (100% by construction at the 3,000 used throughout), and 19 of its 25 misses at 750 characters are rows whose answer lies outside that window. With the end of the row's own thought prepended so that the window is filled to 3,000 characters with the model's reasoning, which names a competing candidate on 1,235 of the 4,368 padded rows, it returns the verified answer on 99.95% (99.92% with a competing candidate). Observationally, on the 36,884 rows where the strict rule parses, agreement with it is lower on longer answer channels, most of all where the window cuts: 99.86% on the 27,707 rows whose answer channel fits in the window, 99.50% on the 8,932 longer rows whose parsed box lies inside it, and 84.1% on the 245 whose parsed box lies before it (second judge 99.92, 99.63 and 84.9%). Of the 124 disagreements, only the judge's reading matches the gold answer in 43 and only the parse's in 42, so where the two disagree neither is more often right. We have not evaluated the judge on an external answer-extraction set such as the Key Answer Finder (KAF) set of Yu et al. (2025); all of its validation uses the responses it scores or rewrites of them.

E.7 The robust-boxed rule

It recovers 59 of the 62 closed gpt-oss NATIVE GPQA rows that the strict rule missed at step 200 (56 with a `\text{}` wrapper, as in `\boxed{\text{C}}`, and three padded with thin spaces), raising the gpt-oss student's GPQA accuracy from 59.5 to 65.1%. It is not a superset of the strict rule: it reads only the last balanced box, so it loses answers that the strict rule finds in an earlier box (in the gold-box step-200 NATIVE GPQA cell it loses eight correct rows and gains one). It changes no original-program contract contrast by more than 2.6 pp; the largest changes from the strict to the robust-boxed rule are from $+2.59$ to $+5.17$ pp for a single checkpoint (gold box, step 25, AIME, a change of three rows in 116), from $+4.31$ to $+2.59$ pp among the seed-42 interactions (gpt-oss AIME) and from -9.14 to -10.15 pp in Table 3 (gold box, GPQA).

F Per-cell table and strict-score decomposition

F.1 Headline contrasts

Table 10 lists the numbers plotted in Figure 4: for each headline contrast, the strict and committed estimates with their item-cluster 95% CIs, and the paired strict-minus-committed artifact with its CI. These CIs are not adjusted for multiplicity. With bootstrap max-$|t|$ simultaneous 95% intervals over each estimand's 14 headline contrasts (exploratory, post hoc; a Bonferroni correction over the 14 gives the same verdicts), every result that the main text's prose calls significant remains significant (also under one max-$|t|$ over all 42 estimates) and none called not significant becomes so; 27 of the 33 unadjusted exclusions of zero survive, the six that do not being Gemini AIME, interaction (committed -9.05, artifact -9.05); gpt-oss GPQA, step 25 (strict -3.87); gpt-oss GPQA, interaction (strict $+3.55$); gold box, PRECLOSED GPQA (strict $+5.46$); and heuristic box, NATIVE AIME (committed -7.76). Two of these six are bold (unadjusted) in Table 1: the Gemini AIME interaction under delivered scoring and the gpt-oss early cost under strict scoring. Figure 4 also shows three contrasts outside this family, with unadjusted intervals: the gpt-oss AIME interaction (committed $+4.31$ $[-5.60, +13.79]$, artifact -1.72 $[-7.33, +3.45]$) and the heuristic box's ΔR on AIME (committed -11.21 $[-25.00, -0.86]$, artifact -12.93 $[-27.59, +0.86]$) and GPQA (committed -0.13 $[-5.08, +4.82]$, artifact -3.43 $[-6.60, -0.25]$).

F.2 Where the strict loss lived

Figure 10 dissects the seed-42 Gemini checkpoint at step 200. On GPQA the PRECLOSED arm commits a visible answer on 99.1% of rows and the NATIVE arm on 97.7% (a further 1.5% end inside an unclosed thought), but 16.2% of the PRECLOSED rows commit without a strict parse—in

Table 10: Strict versus committed readings of the headline contrasts (PRECLOSED−NATIVE unless stated; pp; brackets: unadjusted item-cluster 95% CIs from one bootstrap realization, differing from Table 3's frozen realization by at most 0.12 pp) and the paired strict-minus-committed artifact. These committed analyses are post hoc. Strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; replication: mean of two fresh training seeds; seed 42: the original trajectory; interaction: the contrast at step 200 minus that at step 25; gold box, heuristic box: surgeries appending a gold or a heuristically restated `\boxed{}` answer to the Gemini targets, reported minus the Gemini recipe; ΔR: their change in contract reliance R = NATIVE−PRECLOSED (negative: a smaller contract effect); Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model); pp: percentage points. Step 25: the contrast at step 25; fresh-seed replication: the two fresh training seeds; shaded: the artifact; AIME: AIME 2025 (29 items); GPQA: GPQA Diamond (197 items).

	strict	committed	artifact (strict − committed)
contrast	estimate [95% CI]	estimate [95% CI]	estimate [95% CI]
step-200 contract effects			
Gemini GPQA, replication	−14.34 [−17.51, −11.29]	−0.95 [−3.36, +1.40]	−13.39 [−15.67, −11.29]
Gemini AIME, replication	−15.09 [−23.71, −6.90]	−0.86 [−6.03, +3.88]	−14.23 [−22.84, −5.60]
Gemini GPQA, seed 42	−12.82 [−16.88, −8.88]	−0.89 [−4.44, +2.66]	−11.93 [−14.59, −9.39]
Gemini AIME, seed 42	−23.28 [−34.48, −12.93]	−6.03 [−14.66, +1.72]	−17.24 [−29.31, −6.03]
fresh-seed replication: step-25 effects and pre-registered interactions (step 200 − step 25)			
Gemini GPQA, step 25	−0.19 [−2.66, +2.35]	+16.62 [+13.45, +19.80]	−16.81 [−20.24, −13.45]
Gemini GPQA, interaction	−14.15 [−18.46, −9.96]	−17.58 [−21.76, −13.39]	+3.43 [−0.25, +7.04]
Gemini AIME, interaction	−18.10 [−29.31, −7.76]	−9.05 [−17.24, −1.72]	−9.05 [−18.10, −0.43]
gpt-oss GPQA, step 25	−3.87 [−6.73, −1.08]	−9.58 [−12.50, −6.66]	+5.71 [+2.98, +8.50]
gpt-oss GPQA, interaction	+3.55 [+0.06, +6.98]	+9.45 [+5.77, +13.07]	−5.90 [−9.07, −2.73]
box-appending retrains minus Gemini at step 200 (ΔR uses R = NATIVE−PRECLOSED)			
gold box, PRECLOSED AIME	+31.03 [+18.97, +43.97]	−3.45 [−12.07, +4.31]	+34.48 [+20.69, +49.14]
gold box, PRECLOSED GPQA	+5.46 [+1.65, +9.26]	−5.84 [−9.26, −2.41]	+11.29 [+8.88, +13.83]
gold box, ΔR AIME	−17.24 [−25.86, −9.48]	+0.86 [−7.76, +11.21]	−18.10 [−30.17, −7.76]
gold box, ΔR GPQA	−9.14 [−14.09, −4.19]	+1.90 [−2.92, +6.85]	−11.04 [−13.96, −8.25]
heuristic box, NATIVE AIME	−19.83 [−31.03, −9.48]	−7.76 [−15.52, −0.86]	−12.07 [−19.83, −4.31]

prose, or in a boxed form the rule rejects—and 11.3% of the PRECLOSED rows are so committed *and correct*. The NATIVE and PRECLOSED delivered-answer accuracies are 63.1 and 62.2%. On AIME both arms hit the budget on 41.4% of rows and commit on about 58%; the PRECLOSED arm boxes only 10.3% of rows but is prose-committed and correct on 34.5%, so a strict collapse from 26.7 to 3.4% is 44.0 versus 37.9% committed.

F.3 Consequences for decisions

Across the 96 cells of the Qwen families (excluding the science recipe), strict and delivered accuracy differ by at least 10 pp in 56. At step 25 strict exceeds committed by a median of 21 pp (credit for unfinished and unclosed thoughts); from step 75 on, committed exceeds strict by a median of 5.6 pp and by up to 35 pp (a fresh-seed Gemini checkpoint at step 200, PRECLOSED AIME: 12.1 strict, 47.4% committed). On the seed-42 grid the best checkpoint changes in 3 of the 8 recipe×task×contract panels: under the PRECLOSED contract the strict grader ranks the step-25 Gemini checkpoint best on both tasks—SFT appears to hurt—whereas delivered scoring ranks steps 125 and 200 best.

F.4 Sensitivity to the commitment reading

Reading unclosed NATIVE thoughts on their whole output instead of scoring them uncommitted changes delivered accuracy only where unclosed rows exist: step-25 GPQA NATIVE cells of the Gemini recipe and both surgeries rise by 10–16 pp (e.g. seed 42: 21.3 → 37.1%) and step-25 AIME NATIVE cells by 3–8 pp; step-75 GPQA cells by 3–5 pp; step-125 GPQA cells by 1–3 pp; step-200 cells by at most 1.4 pp; the step-75 and step-125 AIME cells by under 2 pp. No PRECLOSED cell changes, no gpt-oss or science-recipe cell by more than 0.9 pp, and no surgery contrast by more than 0.9 pp. The pre-registered Gemini GPQA interaction is −17.58 pp under the visible-channel reading

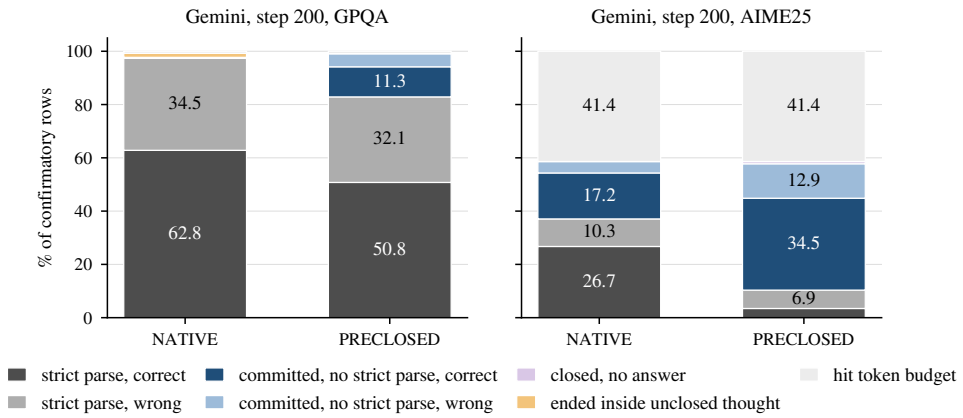

Figure 10: Row-level anatomy of the seed-42 Gemini checkpoint at step 200. The strict grader counts the dark grey segment as correct (plus, on NATIVE GPQA, 0.8 points read out of the 1.5% of rows that ended inside an unclosed thought); the blue segments are answers committed without a strict parse that it counts as wrong. Closed, no answer: closed rows on which the judge returns NONE; AIME25: AIME 2025; GPQA: GPQA Diamond; segments below 6 points are unlabelled. Strict parse: an answer the last-`\boxed{}` rule reads; committed: the final answer the visible answer channel states, read by a gold-blind judge; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; confirmatory rows: the pre-registered rows (each task's first item excluded); Gemini: the recipe distilled from a Gemini-family teacher, named for its teacher.

and -5.33 pp under the whole-output reading, which we used first; its step-25 end is $+16.62$ and $+3.24$ pp and its step-200 end -0.95 and -2.09 pp, moving more than any other step-200 contrast.

F.5 THE LEARNING-CURVE DIP ON FRESH SEEDS (POST HOC)

The seed-42 dip (strict NATIVE AIME accuracy -6.9 points from step 25 to 75, committed $+28.4$) rests on one training seed. We therefore decoded steps 50, 75 and 100 of the two fresh Gemini seeds (AIME 2025, NATIVE, four decoding seeds, T=1, 32,768 tokens, registered judge; exploratory). The prediction, strict step-75 minus step-25 accuracy below zero and committed above zero, pooled over the two seeds, was frozen after the seed-42 dip was known and before any fresh-seed output at these steps existed (sha256 `8cceb25a`…). The first comparison used the registered step-25 runs, decoded with a different engine configuration (tensor parallelism 8, eager mode, 16 concurrent sequences) from the later steps (tensor parallelism 4, CUDA graphs, 128 concurrent sequences). Before re-running step 25 with the later steps' configuration we designated that same-configuration comparison as primary and the mixed one as secondary (sha256 `99f3845d`…). Table 13 gives the cells and Table 14 the contrasts. The prediction holds in both analyses, but the strict dip excludes zero only in the primary one: the engine configuration alone moves step-25 strict accuracy by 4.3 and 6.0 points (delivered accuracy by at most 2.6), about the size of the dip. Delivered accuracy rises by 31.0 to 33.6 points in every comparison and seed.

G TERMINATION OUTCOMES AND AN UNINFORMATIVE FROZEN READOUT

The frozen analyzers reported `normal_closure` as "returned text ends with the literal end-of-turn string". The serving stack does not include the stop token in the returned text, so this readout was ≈ 0 in every cell and the earlier statement "closure deltas are exactly zero" carried no information. Auditing all 51,072 rows of the 28 evaluated checkpoints: the finish reason is `stop` for 40,384 rows and `length` for 10,688; every `length` row has exactly 32,768 generated tokens; every `stop` row's last token is the end-of-turn token. The finish reason is therefore exhaustively {end-of-turn, budget}. The channel state at termination is a further dimension the frozen analyzers did not record: among the 20,055 confirmatory NATIVE end-of-turn rows, 1,164 (5.8%) contain no generated `</think>`, i.e. the model emitted end-of-turn inside its thought (§4); three quarters

Table 11: Exact allocation of strict-score differences (pp) to state mix, parse within state and correctness given parse (Shapley over the three factors of equation 5, per state, summed). Convention: a factor that is undefined in one cell (π_s when $P(s) = 0$; ρ_s when $P(s) = 0$ or $\pi_s = 0$) takes the other cell's value, so that a state or a parse that appears or disappears is charged to mix or parse, never to an unobserved factor. "sens." is the largest change in any component under the alternative convention that borrows only when $P(s) = 0$ and sets an undefined ρ_s to 0. The two conventions differ only where a state has rows but no parse in one cell of a contrast and parsed rows in the other (8 state–cell combinations among the compared cells have rows but no parse); 10 of the 30 contrasts are affected, by up to 8.4 pp in the reliability term (Gemini AIME recovery). Among the bars of Figure 9 and the terms quoted in Appendix A.5, only the Gemini recoveries and the AIME collapses move by more than 1.2 pp; the Gemini AIME collapse becomes -75.1 mix, $+7.8$ parse and $+9.5$ reliability. No convention changes a state's total contribution, only its split among the three factors. Both conventions borrow when $P(s) = 0$, so a state that appears or disappears counts as mix; imputing 0 instead would split its contribution equally among the three factors (largest change: Gemini GPQA collapse, -40.7 mix instead of -16.3). Last column: the same difference under visible-channel delivered scoring. Seed-42 trajectories and the base anchor; NATIVE unless the contrast says otherwise. Allocations are descriptive, not causal. Each entry is rounded once from its exact value, so a row's three terms, or its endpoints, can differ from its Δstrict by 0.1. Contrasts: collapse (base→step 25), dip (step 25→75), recovery (step 25→200), net (base→step 200), contract (NATIVE→PRECLOSED), recipe gap (Gemini→gpt-oss); NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; strict: accuracy under the last-`\boxed{}` rule at the two endpoints; committed: the final answer the visible answer channel states, read by a gold-blind judge; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model); pp: percentage points. Base: the untouched model.

			allocation of Δstrict				
contrast	strict	Δstrict	mix	parse	reliability	sens.	Δcommitted
GPQA							
base contract: NATIVE → PRECLOSED	71.7→72.1	+0.4	+0.0	+0.0	+0.4	0.0	+0.4
Gemini contract at step 200: NATIVE → PRECLOSED	63.6→50.8	−12.8	+0.0	−9.9	−3.0	0.0	−0.9
Gemini contract at step 25: NATIVE → PRECLOSED	57.5→57.0	−0.5	+4.6	−3.3	−1.8	0.0	+17.6
Gemini dip: step 25 → 75	57.5→60.3	+2.8	+21.6	−10.2	−8.6	0.0	+35.7
Gemini net: base → step 200	71.7→63.6	−8.1	−0.8	−0.3	−7.1	0.0	−8.6
Gemini recovery: step 25 → 200	57.5→63.6	+6.1	+25.2	−11.1	−8.1	7.8	+41.8
Gemini collapse: base → step 25	71.7→57.5	−14.2	−16.3	−0.9	+3.0	0.0	−50.4
gpt-oss contract at step 200: NATIVE → PRECLOSED	59.5→58.8	−0.8	+0.4	−0.9	−0.2	0.0	+0.5
gpt-oss contract at step 25: NATIVE → PRECLOSED	57.1→55.3	−1.8	−3.1	−0.4	+1.7	0.0	−9.3
gpt-oss dip: step 25 → 75	57.1→49.9	−7.2	+6.8	−22.3	+8.3	0.0	+18.9
gpt-oss net: base → step 200	71.7→59.5	−12.2	−8.6	−6.2	+2.6	0.0	−7.0
gpt-oss recovery: step 25 → 200	57.1→59.5	+2.4	+10.7	−13.4	+5.1	0.0	+21.8
gpt-oss collapse: base → step 25	71.7→57.1	−14.6	−11.5	−0.8	−2.3	0.0	−28.8
recipe gap at step 200 NATIVE: Gemini → gpt-oss	63.6→59.5	−4.1	−7.7	−5.2	+8.9	0.2	+1.6
recipe gap at step 200 PRECLOSED: Gemini → gpt-oss	50.8→58.8	+8.0	−6.9	+3.9	+11.0	0.0	+3.0
AIME 2025							
base contract: NATIVE → PRECLOSED	81.9→77.6	−4.3	−2.4	+0.8	−2.7	0.0	−4.3
Gemini contract at step 200: NATIVE → PRECLOSED	26.7→3.4	−23.3	+0.0	−14.1	−9.2	0.0	−6.0
Gemini contract at step 25: NATIVE → PRECLOSED	24.1→27.6	+3.4	+2.3	+0.2	+0.9	0.0	+6.0
Gemini dip: step 25 → 75	24.1→17.2	−6.9	+21.6	−16.8	−11.7	1.0	+28.4
Gemini net: base → step 200	81.9→26.7	−55.2	−19.5	−22.0	−13.7	0.0	−37.9
Gemini recovery: step 25 → 200	24.1→26.7	+2.6	+32.6	−23.7	−6.3	8.4	+43.1
Gemini collapse: base → step 25	81.9→24.1	−57.8	−72.1	+12.2	+2.2	7.4	−81.0
gpt-oss contract at step 200: NATIVE → PRECLOSED	68.1→69.0	+0.9	+2.3	+1.8	−3.3	0.0	+0.0
gpt-oss contract at step 25: NATIVE → PRECLOSED	26.7→23.3	−3.4	−1.9	−2.1	+0.5	0.0	−2.6
gpt-oss dip: step 25 → 75	26.7→67.2	+40.5	+38.4	−3.2	+5.4	5.4	+56.9
gpt-oss net: base → step 200	81.9→68.1	−13.8	−15.3	−0.5	+1.9	0.4	−14.7
gpt-oss recovery: step 25 → 200	26.7→68.1	+41.4	+35.2	−7.1	+13.3	0.0	+54.3
gpt-oss collapse: base → step 25	81.9→26.7	−55.2	−49.2	−0.2	−5.7	5.0	−69.0
recipe gap at step 200 NATIVE: Gemini → gpt-oss	26.7→68.1	+41.4	+8.1	+19.9	+13.3	0.5	+23.3
recipe gap at step 200 PRECLOSED: Gemini → gpt-oss	3.4→69.0	+65.5	+5.5	+36.6	+23.5	1.2	+29.3

are step-25 rows of the Gemini recipe and its two surgeries, and for gpt-oss unclosed rows are at most 1.7% of any cell. Under PRECLOSED the closer is part of the prompt and the question does not arise. Re-emission of `<think>` anywhere in the output occurs in 0.97% (NATIVE) and 1.07%

Table 12: Selected cells (%): strict accuracy, visible-channel delivered accuracy, strict parse rate, commitment rate, budget-hit rate, share of NATIVE rows that ended inside an unclosed thought, and share of rows committed without a strict parse (N = NATIVE, P = PRECLOSED; training seed 42 unless given). All 112 cells, under both commitment readings, are in the supplementary material. Strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; heuristic box, gold box: Gemini targets with a heuristic or gold `\boxed{}` answer appended; commitment readings: visible channel only, or unclosed rows read on their whole output; step: optimizer step; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model). Task/arm: task and contract (N, P); seeds 271828 and 314159: the fresh training seeds; AIME: AIME 2025; GPQA: GPQA Diamond.

			accuracy		share of rows				
recipe	step	task/arm	strict	committed	parse	commit	budget	unclosed	unparsed-committed
Gemini	25	AIME N	24.1	0.9	40.5	0.9	95.7	3.4	0.0
Gemini	25	AIME P	27.6	6.9	44.0	8.6	90.5	0.0	2.6
Gemini	25	GPQA N	57.5	21.3	88.7	27.9	47.7	24.4	0.5
Gemini	25	GPQA P	57.0	39.0	87.1	54.8	44.4	0.0	3.9
Gemini	200	AIME N	26.7	44.0	37.1	58.6	41.4	0.0	21.6
Gemini	200	AIME P	3.4	37.9	10.3	57.8	41.4	0.0	47.4
Gemini	200	GPQA N	63.6	63.1	98.6	97.7	0.8	1.5	0.4
Gemini	200	GPQA P	50.8	62.2	82.9	99.1	0.9	0.0	16.2
gpt-oss	25	AIME N	26.7	12.9	48.3	16.4	79.3	1.7	0.9
gpt-oss	25	AIME P	23.3	10.3	42.2	12.9	85.3	0.0	1.7
gpt-oss	25	GPQA N	57.1	42.9	92.5	62.2	36.7	0.9	0.6
gpt-oss	25	GPQA P	55.3	33.6	90.7	49.6	49.6	0.0	1.6
gpt-oss	200	AIME N	68.1	67.2	69.8	69.0	29.3	0.0	0.0
gpt-oss	200	AIME P	69.0	67.2	74.1	73.3	26.7	0.0	0.9
gpt-oss	200	GPQA N	59.5	64.7	80.1	86.8	13.2	0.0	7.9
gpt-oss	200	GPQA P	58.8	65.2	79.1	87.4	12.6	0.0	9.0
heuristic box	200	AIME N	6.9	36.2	12.9	53.4	45.7	0.0	40.5
heuristic box	200	AIME P	7.8	41.4	11.2	63.8	36.2	0.0	52.6
gold box	200	AIME N	40.5	41.4	55.2	55.2	43.1	0.9	0.9
gold box	200	AIME P	34.5	34.5	51.7	52.6	47.4	0.0	0.9
gold box	200	GPQA N	59.9	59.1	97.3	96.3	1.5	2.0	1.1
gold box	200	GPQA P	56.2	56.3	97.3	98.2	1.5	0.0	1.0
Gemini, seed 271828	25	GPQA N	59.4	24.1	87.9	29.4	49.9	20.7	0.0
Gemini, seed 271828	25	GPQA P	58.1	41.1	87.6	57.4	42.4	0.0	3.8
Gemini, seed 271828	200	GPQA N	62.4	61.3	98.7	97.3	0.8	1.9	0.1
Gemini, seed 271828	200	GPQA P	50.5	62.2	82.4	99.7	0.3	0.0	17.4
Gemini, seed 314159	25	GPQA N	57.5	22.0	89.6	27.0	50.5	22.5	0.3
Gemini, seed 314159	25	GPQA P	58.4	38.2	87.4	55.1	44.7	0.0	3.6
Gemini, seed 314159	200	GPQA N	63.1	62.6	98.1	97.5	1.1	1.4	0.6
Gemini, seed 314159	200	GPQA P	46.3	59.8	79.3	98.7	1.3	0.0	19.8

Table 13: The fresh-seed dip: Gemini recipe, AIME 2025, NATIVE (%; 116 rows per cell, 29 items $\times$ 4 decoding seeds). 25 (registered run): the registered step-25 decoding; 25 (re-run): step 25 decoded with the configuration of steps 50–100. Budget, unclosed: termination states. Exploratory, post hoc.

training seed	step	strict	committed	budget	unclosed
271828	25 (registered run)	25.9	1.7	87.9	10.3
271828	25 (re-run)	31.9	4.3	86.2	9.5
271828	50	17.2	28.5	60.3	5.2
271828	75	16.4	35.3	57.8	0.9
271828	100	26.7	39.7	51.7	0.9
314159	25 (registered run)	22.4	0.9	91.4	7.8
314159	25 (re-run)	26.7	0.9	92.2	6.9
314159	50	21.6	29.3	61.2	2.6
314159	75	24.1	32.8	61.2	0.0
314159	100	21.6	34.5	53.5	0.0

(PRECLOSED) of rows, concentrated at step 25 and at most 0.13% of any step-200 cell. No strict accuracy or censoring number is affected.

Table 14: The fresh-seed dip: step 75 minus step 25 (pp) with item-cluster 95% CIs. One configuration: step 25 re-run with the configuration of steps 50–100 (primary, designated before the re-run); mixed: the registered step-25 decoding (secondary).

engine configuration	seeds	strict	committed
one configuration (primary)	both seeds	-9.1 $[-17.7, -0.9]$	$+31.5$ $[+18.5, +44.8]$
	seed 271828	-15.5 $[-25.0, -6.0]$	$+31.0$ $[+18.1, +44.0]$
	seed 314159	-2.6 $[-12.9, +7.8]$	$+31.9$ $[+18.1, +45.7]$
mixed (secondary)	both seeds	-3.9 $[-10.3, +2.6]$	$+32.8$ $[+19.8, +46.1]$
	seed 271828	-9.5 $[-18.1, -0.9]$	$+33.6$ $[+20.7, +47.4]$
	seed 314159	$+1.7$ $[-3.5, +7.8]$	$+31.9$ $[+18.1, +45.7]$

H PROOF OF THE BOUND AND ADDITIONAL ANALYSES

Table 15: Two clocks along the seed-42 Gemini trajectory (%, GPQA unless noted; —: the base model has no non-answer rows). The credit terms of equation 1 peak at step 25 and fall to a few points by step 75; the unparsed-answer term of equation 2 appears by step 75 and stays. Two clocks: the credit for boxed guesses in non-answer (budget-hit or unclosed) rows fades as termination recovers, while unparsed prose answers grow; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; strict parse: an answer the last-`\boxed{}` rule reads; committed: the final answer the visible answer channel states, read by a gold-blind judge; Gemini: the recipe distilled from a Gemini-family teacher, named for its teacher. Base: the untouched model; non-answer rows: budget-hit or unclosed rows.

		Gemini, step			
	base	25	75	125	200
median output tokens, NATIVE	7,352	30,174	17,086	15,390	15,296
budget hits, NATIVE	0.0	47.7	3.7	2.5	0.8
ended inside unclosed thought, NATIVE	0.0	24.4	7.7	2.5	1.5
non-answer rows with a strict parse, NATIVE	—	85.0	52.2	42.5	55.6
closed rows without a strict parse, PRECLOSED	0.0	8.2	24.8	24.8	16.4
closed rows without a strict parse, PRECLOSED AIME	0.0	36.4	78.8	78.9	82.4
strict accuracy, NATIVE	71.7	57.5	60.3	61.3	63.6
delivered accuracy, NATIVE	71.7	21.3	57.0	60.3	63.1

H.1 STATEMENT AND PROOF OF THE BOUND

Let every row carry a termination state $s \in \{\mathrm{B}, \mathrm{U}, \mathrm{C}\}$, a parse flag π with parsed answer y_π when π holds, and, on rows in state C, a delivered answer y_c (possibly NONE). Strict scores a row correct iff π and $y_\pi = g$; delivered iff $s = \mathrm{C}$ and $y_c = g$. Subtracting equation 2 from equation 1 and splitting the closed, parsed rows by whether $y_\pi = y_c$,

$$\begin{aligned}\mathrm{Strict} - \mathrm{Delivered} = {} & P(\pi, \mathrm{B} \cup \mathrm{U}, y_\pi{=}g) - P(\mathrm{C}, \neg\pi, y_c{=}g) \\ & + P(\mathrm{C}, \pi, y_\pi{\neq}y_c, y_\pi{=}g) - P(\mathrm{C}, \pi, y_\pi{\neq}y_c, y_c{=}g).\end{aligned}$$

The first and third terms are non-negative and bounded by $a = P(\pi, \mathrm{B} \cup \mathrm{U})$ and δ; the second and fourth by $b = P(\mathrm{C}, \neg\pi)$ and δ, where a and b are the two judge-free counts of equation 3. Dropping the negative terms gives the upper bound of equation 3, dropping the positive ones the lower bound. The upper bound is attained when every guess read out of a non-answer state is correct, no unparsed answer is, and every disagreement row is one the parser scores correct and the judge wrong; the lower when the reverse holds. In δ a judge verdict of NONE or an out-of-domain reading on a parsed row counts as a disagreement. On the Qwen bed the conditional rate $P(y_\pi{\neq}y_c \mid \mathrm{C}, \pi)$ is 0.30% (108 of 35,626 rows; NONE alone 0.02%; the highest cell rate is 2 of 19 rows, 10.5%) and the unconditional mass δ of equation 3 is 0.21% of the 50,624 confirmatory rows. Equation 5 follows from equation 1 by conditioning each term on its state. Three forms of the bound should be kept apart. With δ, equation 3 holds by construction, so evaluating it checks the computation, not the data: neither side is violated in the 32 seed-42 cells or the 4 base cells, evaluated exactly on every row, or in any of the 112 cells, reconstructed from cell statistics. Without δ, the symmetric form $|\mathrm{Strict} -$

$\text{Delivered}| \leq a + b$, which is the form registered as P1 (Appendix J), holds with zero tolerance in all 112 Qwen cells and all 60 Gemma cells and fails in one held-out cell. Without δ the *directional* form $-b \leq \text{Strict} - \text{Delivered} \leq a$ is exceeded in two of the 112 cells (two science-recipe GPQA PRECLOSED cells with $a = 0$, each by one disagreement row of 788, $+0.13$ pp) and in that same held-out cell: the masses need no judge, the guarantee needs δ. Of the 112 cells, 52 have a positive artifact (median ratio to the upper bound of equation 3, $a + \delta$, 0.50), 45 a negative one (median ratio to its lower bound, $-b-\delta$, 0.66) and 15 a zero one; the larger judge-free count gives the sign of 91 of the 97 nonzero artifacts, and the six exceptions are positive artifacts of at most 0.89 pp with $b > a$.

H.2 LENGTH RELATIVE TO THE TRAINING DISTRIBUTION

Training dropped targets longer than 32,768 tokens (5.33% of Gemini and 16.16% of gpt-oss targets), so no trained target reaches the evaluation budget. Among all targets (the teacher-census sample of Appendix B), whole Gemini targets, thought plus answer, have a 90th percentile of 28,034 tokens and whole gpt-oss targets 44,041; the trained (kept) distribution is bounded by 32,768 by construction. At step 25, on the six Gemini and gpt-oss trajectories and both contracts (Table 16), 35.0–52.9% of the students' GPQA outputs and 76.7–95.7% of their AIME outputs sit at that bound, and 39.0–57.9% of their GPQA outputs exceed the whole Gemini targets' 90th percentile, against 10% by definition for the Gemini teacher and 0.0–0.1% for the base model.

Table 16: Termination at step 25 on the six Gemini and gpt-oss trajectories (two recipes × three training seeds; N = NATIVE, P = PRECLOSED; confirmatory rows, 32,768-token budget): median GPQA output length (tokens) and GPQA and AIME budget-hit rates (%), with the AIME budget-hit rate at step 200 for comparison. The base model's GPQA medians are 7,352 (N) and 7,507 (P) tokens with no budget hit; its AIME budget-hit rates are 12.9 (N) and 15.5% (P). Budget hit: the output reaches the 32,768-token limit; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model).

		step 25						step 200	
		GPQA median		GPQA budget		AIME budget		AIME budget	
recipe	seed	N	P	N	P	N	P	N	P
Gemini	42	30,174	27,710	47.7	44.4	95.7	90.5	41.4	41.4
Gemini	271828	32,747	25,624	49.9	42.4	87.9	88.8	42.2	40.5
Gemini	314159	32,768	27,960	50.5	44.7	91.4	88.8	38.8	32.8
gpt-oss	42	19,392	32,054	36.7	49.6	79.3	85.3	29.3	26.7
gpt-oss	271828	20,768	32,768	39.5	50.3	77.6	81.9	24.1	25.0
gpt-oss	314159	19,111	32,768	35.0	52.9	76.7	82.8	25.0	24.1

H.3 WHAT PRECEDES A RE-EMITTED CLOSER

Under PRECLOSED, the base model and the gpt-oss student re-emit `</think>` on almost every row (Table 5). When the Gemini student at step 200 does not re-emit it, the last 200 characters of its GPQA output carry an answer signature ("Final Answer", a box, or "Answer: X") on 85.2% of rows; when it does, the text before the closer carries one on 30.6%. The gpt-oss student closes after an answer signature on 87.9% of rows, then boxes again, which is its teacher's pattern (box inside the thought on 98.1% of targets, in the answer on 100%). The Gemini student's answer-channel derivation therefore ends the way an answer segment ends—state the answer, stop—while the gpt-oss student's ends the way a thought ends.

H.4 PROSE COMMITMENTS ARE NOT DRIVEN BY NUMERIC OPTIONS

On GPQA, 127 of the 197 confirmatory items have a numeric correct option. Among the Gemini PRECLOSED committed rows from step 75 on, the unparsed-commitment rate is 19.7% on numeric-gold items and 22.3% on textual-gold items. Numeric items differ in another way: in 15.7% of their unparsed rows the last 600 characters of the answer channel name no option letter in an answer phrase (such as "answer is X" or "go with X") or as "(X)", against 4.8% for textual items. The judge never sees the option texts; Appendix E reports the answers it returns that do not occur literally in the judged answer-channel tail.

H.5 ANSWER RELIABILITY BY TERMINATION STATE

On the NATIVE arm at step 25, take the answer a whole-output strict reading would credit: the box after the closer in a closed row, the last box in the thought otherwise. On GPQA that answer is correct on 76.4–81.9% of closed rows, 65.4–66.5% of unclosed rows and 52.4–58.1% of budget-hit rows across the three Gemini training seeds (68.9, 57.1 and 47.6% for the gpt-oss student; 74.6 and 55.6% for closed and budget-hit rows of the gpt-oss student at step 200). The judge's commitment on unclosed rows is correct on 60.5–65.6%. Part of this gradient is item difficulty: the items whose Gemini row closed at step 25 are ones the base model answers correctly 81–83% of the time, those whose row hit the budget 66–69%. Comparing each boxed row with the base model's accuracy on its own item (so both sides condition on the parsed rows), closed-row answers are within -4.5 to $+0.8$ pp of the base and unclosed-row answers within -6.1 to -1.7 pp, whereas the boxed guesses in truncated thoughts are 11.0–15.4 pp below it (-6.9, -14.3 and -16.9 for the gpt-oss student). This is a descriptive comparison with the base on associated items, not the counterfactual of the same student finishing. Pooled over budget-hit and unclosed rows, the boxes that the strict rule reads in the 24 step-25 cells (three training seeds, two recipes, two tasks, two contracts) are correct on 42–61% (AIME: 42–61%, against a chance rate of 0.1% for an integer from 0 to 999; GPQA: 48–61%, against 25%). Unfinished rows are far from empty, but a grader that pools the three states under one number moves with the state mix even when the reliability within each state is constant.

H.6 A SYMBOLIC-EQUIVALENCE CHECKER CAN REPAIR FORMAT, NOT TERMINATION

Mathematics harnesses often replace the regular expression with a computer-algebra equivalence checker (Gao et al., 2023). We re-scored the 16,272 confirmatory rows of the seed-42 trajectories and the base anchor with math-verify 0.9.0 (Kydlíček, 2025; exploratory, post hoc; no parse or verification timed out). It reads the segment the strict rule reads, cut to start 300 characters before the segment's last `\boxed` or, without one, to its last 3,000 characters; in 119 rows the box the strict rule scores lies before that window. It repairs most of the *format* term where answers are integers: on AIME the Gemini step-200 contract effect is -23.3 pp strict, -7.8 pp under math-verify and -6.0 pp committed (the PRECLOSED cell scores 3.4, 36.2 and 37.9%), because it reads "Result is 167." and boxed equations; on the 228 contested AIME rows that closed and that the judge answered, it is correct on 69.7% against the judge's 75.4%. On GPQA the repair depends on how option letters are extracted. As run, with the gold letter read as a string and the output read by the LaTeX extractor before the letter strings, it repairs less (-12.8, -5.8 and -0.9 pp; 21.6% against 69.3% on the 1,199 such GPQA rows): its letter strings do not match "(C)", and a boxed `\text{C}` parses to a symbol, which never equals a letter read as a string (of the 561 such rows that contain one, it credits none and the judge 396). Under the configuration of verl's `math_verify` scorer (see the census below), applied to the same window, it reads boxed letters but not "Answer: C" and repairs almost none of the step 200 effect (-12.7 pp; 36.0% on those rows); with letter wrappers such as "(C)" rewritten as bare letters and only letter strings extracted, it repairs most of it (-2.3 pp; 66.9%). It does nothing for the *state* terms, since any checker that reads the whole output credits what it finds in unfinished thoughts: in the eight step-25 cells math-verify exceeds delivered accuracy by 11.2–39.5 pp (strict rule: 12.9–36.2), the Gemini step-25 GPQA contract contrast is -0.6 pp against $+17.6$ pp committed, and the gpt-oss recipe's early cost of pre-closing is estimated at -2.0 pp against -9.3 pp committed, a size that itself depends on how truncated answers are counted (Appendix I.7). Neither alternative GPQA configuration changes this: in the four step-25 GPQA cells math-verify still exceeds delivered accuracy by at least 13.5 pp, and the Gemini and gpt-oss contrasts are -0.6 and -2.7 pp under the verl configuration and $+0.6$ and -1.4 pp under the letter configuration. Equivalence checking addresses what counts as the same answer; the identity is about where an answer is read from.

H.7 HARNESS CENSUS

We read the answer-extraction code of four public reasoning-model harnesses at snapshots dated December 2025 to February 2026. NeMo-Skills grades the whole generation, thoughts included, unless `parse_reasoning=True` is set (default `False`); with it set, the graded text is what follows the last `</think>` and an unclosed thought is graded as empty—the visible-channel rule of §4. NeMo-RL's math environment and verl's `math_dapo` and `math_verify` scorers pass the whole response to the extractor and take the last box wherever it occurs; ms-swift's `MathORM`

takes the *first* `\boxed` in the text. None of the four separates budget-hit from terminated rows in its score. The whole-output reading that produces truncation and unclosed-thought credit is therefore the default, not an idiosyncrasy of our frozen rule.

H.8 A JUDGE-FREE VERSION OF THE PROTOCOL (EXPLORATORY, POST HOC)

Scoring budget-hit and unclosed rows as unanswered and reading only the text after the thought closer—the last balanced `\boxed{}` with `\text{}` unwrapped, else an answer line, compared by math-verify on AIME and by letter on GPQA (variant (b))—recovers the step-25 contract effect that the strict rule missed (Gemini GPQA: $+15.67$ against $+16.62$ pp committed and -0.19 pp strict) and matches delivered accuracy within 1.72 pp in all 78 of the 112 discovery cells in which u, the share of rows that close without an extractable answer, is below 2% (held-out: 15 of 20 cells, within 0.83 pp). A rate $u \geq 2\%$ flags all 29 cells (26 discovery, 3 held-out) that the pipeline misses by more than 2 pp; all come from the Gemini recipe or its surgeries, whose answers are often prose, and there it under-credits by up to 31.90 pp, so a judge is still needed. The reasoning parser does most of the work: without it, a better checker (robust boxed with math-verify) stays far from delivered accuracy (median 5.53 against 9.87 pp for the strict rule; 59 against 71 cells off by more than 2 pp; Table 17). A variant (c), added post hoc, that falls back to a flexible extraction of the last number or letter shrinks the residual (at most 11.21 pp) but breaks the flag: 8 of its 12 cells off by more than 2 pp have $u < 2\%$.

Table 17: Judge-free grading versus delivered accuracy (exploratory, post hoc). Top: per-cell |score−committed| (pp) over the 112 discovery cells (AIME 2025, GPQA Diamond) and the 20 held-out cells (AIME 2024, GPQA-main). Strict rule: the frozen last-`\boxed{}` rule; closed rows only: strict with budget-hit and unclosed rows scored unanswered (the reasoning parser); robust boxed, no reasoning parser: last balanced box with math-verify on the whole output; (a): gate plus robust boxed; (b): (a) plus answer lines; (c): (b) plus a flexible-extract fallback, added post hoc (†); 90th pct.: 90th percentile; u: share of rows that close without an extractable answer. Bottom: the headline contrasts of Table 10 (PRECLOSED−NATIVE unless stated) under delivered scoring (judge-read visible answer) and judge-free (b), with (b) minus committed and its item-cluster 95% CI; bold: CI excludes 0; ΔR: change in contract reliance R = NATIVE−PRECLOSED; Gemini, gpt-oss: recipes named for their teachers; pp: percentage points.

	original program, 112 cells				held-out panel, 20 cells			
scorer	median	90th pct.	max	> 2 pp	median	90th pct.	max	> 2 pp
strict rule	9.87	29.30	37.18	71	11.60	21.04	41.67	13
strict, closed rows only	0.86	18.97	35.34	42	0.40	11.83	41.67	5
robust boxed, no reasoning parser	5.53	28.28	36.68	59	11.43	20.27	41.67	11
judge-free (a): gate + robust boxed	0.25	17.16	35.34	30	0.10	11.83	41.67	3
judge-free (b): (a) + answer lines	0.13	16.03	31.90	26	0.20	5.27	28.33	3
(c): (b) + flexible fallback†	0.00	2.02	11.21	12	0.10	0.83	2.40	1
(b), cells with $u < 2\%$		78 cells	1.72	0		15 cells	0.83	0

			judge-free (b)		
contrast	committed	strict − committed	contrast	− committed	95% CI
Gemini GPQA, replication	-0.95	$\mathbf{-13.39}$	$\mathbf{-5.71}$	$\mathbf{-4.76}$	$[-6.03, -3.55]$
Gemini AIME, replication	-0.86	$\mathbf{-14.22}$	$\mathbf{-11.21}$	$\mathbf{-10.34}$	$[-18.10, -3.45]$
Gemini GPQA, seed 42	-0.89	$\mathbf{-11.93}$	$\mathbf{-5.08}$	$\mathbf{-4.19}$	$[-5.71, -2.79]$
Gemini AIME, seed 42	-6.03	$\mathbf{-17.24}$	$\mathbf{-18.10}$	$\mathbf{-12.07}$	$[-23.28, -2.59]$
Gemini GPQA, step 25	$\mathbf{+16.62}$	$\mathbf{-16.81}$	$\mathbf{+15.67}$	$\mathbf{-0.95}$	$[-1.59, -0.38]$
Gemini GPQA, interaction	$\mathbf{-17.58}$	$+3.43$	$\mathbf{-21.38}$	$\mathbf{-3.81}$	$[-5.20, -2.47]$
Gemini AIME, interaction	$\mathbf{-9.05}$	$\mathbf{-9.05}$	$\mathbf{-18.53}$	$\mathbf{-9.48}$	$[-16.81, -2.59]$
gpt-oss GPQA, step 25	$\mathbf{-9.58}$	$\mathbf{+5.71}$	$\mathbf{-9.77}$	-0.19	$[-0.63, +0.25]$
gpt-oss GPQA, interaction	$\mathbf{+9.45}$	$\mathbf{-5.90}$	$\mathbf{+9.39}$	-0.06	$[-0.70, +0.57]$
gold box, PRECLOSED AIME	-3.45	$\mathbf{+34.48}$	$\mathbf{+24.14}$	$\mathbf{+27.59}$	$[+16.38, +39.66]$
gold box, PRECLOSED GPQA	$\mathbf{-5.84}$	$\mathbf{+11.29}$	-1.78	$\mathbf{+4.06}$	$[+2.66, +5.58]$
gold box, ΔR AIME	$+0.86$	$\mathbf{-18.10}$	$\mathbf{-11.21}$	$\mathbf{-12.07}$	$[-23.28, -2.59]$
gold box, ΔR GPQA	$+1.90$	$\mathbf{-11.04}$	-3.30	$\mathbf{-5.20}$	$[-6.85, -3.68]$
heuristic box, NATIVE AIME	$\mathbf{-7.76}$	$\mathbf{-12.07}$	$+1.72$	$\mathbf{+9.48}$	$[+1.72, +18.10]$

H.9 HEDGING

Among contested GPQA rows (terminated, no strict parse, judge answered), the last 3,000 characters name two or more different option letters in answer-like phrases on 32–40% of Gemini rows (native 39.5%, PRECLOSED 38.9%; fresh seeds 31.9% and 33.9%) against 6–10% of gpt-oss rows. Prose commitment comes with hedging, and the hedging is a property of the recipe rather than of the contract. The judge resolves it by taking the last commitment; a grader that reports commitment should report this rate as well.

H.10 THE BOUND ON GEMMA-4-31B AND A DECISION TABLE

The symmetric judge-free form of equation 3 holds in all 60 Gemma-4-31B cells (15 runs, two tasks, two contracts). These runs are scored with the strict rule they were originally scored with, which in bare-assistant rows reads the whole output, closed thought included, and with a judge that reads the tail of that same span (Appendix R), so their masses and artifacts are measured under a related rather than identical scoring policy. Re-scored under the Qwen bed's reading (same appendix), the symmetric form again holds in all 60 cells, and the Gemma columns of Table 18 are unchanged except at a threshold of 20 (41 cells). Table 18 gives, for both beds, the number of cells whose judge-free sum falls below a threshold and the largest artifact realised among them; the Qwen regularity "below 2, at most 1.14" becomes "below 2, at most 1.72" on Gemma, so it is an empirical regularity of a bed, not a guarantee.

Table 18: Decision table for the judge-free counts $a + b$ of equation 3 (Gemma cells under the scoring policy described above): cells whose sum lies below a threshold and the largest realised $|\mathrm{strict} - \mathrm{committed}|$ among them (pp). An empirical regularity of each bed, not a guarantee. a: share of rows with a strict parse that hit the budget or ended inside an unclosed thought; b: share of closed rows without a strict parse (neither needs a judge); strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; max artifact: the largest |strict−committed|; bed: the evaluated cells of one base model; pp: percentage points. Threshold: cut-off on $a+b$ (pp); Qwen3 bed: the 112 discovery cells; Gemma-4 bed: the 60 Gemma-4-31B cells.

	Qwen3 bed (112 cells)		Gemma-4 bed (60 cells)	
threshold	cells	max artifact	cells	max artifact
1	12	0.86	2	0.86
2	20	1.14	5	1.72
3	28	2.59	9	2.59
5	39	2.59	14	2.59
10	51	6.47	35	6.47
20	61	12.07	45	15.36

I RE-SCORING WITHOUT NEW GENERATION

The analyses below re-read the stored generations and judgements; none generates or judges anything new (exploratory, post hoc; item-cluster bootstrap, B=10,000).

I.1 TWO MECHANISMS BEHIND THE DECISIONS

Table 19 separates the credit that the strict rule gives to boxes in unfinished thoughts from the way it reads finished answers. Keeping the credit but judging finished answers keeps the step-25 tie and removes the step-200 cost and the rescue; removing the credit but keeping the strict reading of finished answers (the reasoning parser) does the opposite. The dip survives only when the credit is kept and finished answers are read by pattern (strict or regex). The interaction, found by every reader that gives unfinished rows no credit, is not clear when they are credited and finished answers are judged.

I.2 PUBLIC-MODEL RANKINGS

On GPQA-main under NATIVE, five of the twelve models leave 24.8–100.0% of their finished rows without a strict parse (the others at most 2.0%), so their b is large; every one of the 13 strict inversions

Table 19: **Crediting unfinished boxes decides the step-25 comparison; reading finished answers decides the step-200 cost and the rescue; the dip needs both.** Each decision of Table 1 under four readers that separate the two mechanisms (exploratory, post hoc). + unfinished boxes, judge: budget-hit and unclosed rows scored by the strict rule (a correct box is credited), finished rows by the judge; + unfinished boxes, regex: the same with finished rows read by the lenient regex; judge-free (b): the pipeline of Appendix H. Other columns and marks as in Table 1 (bold: 95% CI excludes 0; shaded: verdict differs from strict). Top: original items (fresh training seeds, except the seed-42 rescue and dip); bottom: the held-out panel (seed-42 checkpoints).

decision	task	strict	reasoning parser	lenient regex	delivered	+ unfinished boxes, judge	+ unfinished boxes, regex	judge-free (b)
pre-closing hurts, step 200	GPQA	**−14.3**	**−13.6**	**−2.6**	−0.9	−1.7	**−3.4**	**−5.7**
	AIME	**−15.1**	**−15.1**	**−11.2**	−0.9	−0.9	**−11.2**	**−11.2**
the effect grows, 25→200	GPQA	**−14.2**	**−28.3**	**−18.9**	**−17.6**	−3.4	**−4.7**	**−21.4**
	AIME	**−18.1**	**−22.0**	**−18.5**	**−9.1**	−5.2	**−14.7**	**−18.5**
better contract, step 25	GPQA	−0.2	**+14.7**	**+16.2**	**+16.6**	+1.7	+1.3	**+15.7**
gpt-oss: early cost	GPQA	**−3.9**	**−9.8**	**−9.4**	**−9.6**	**−3.6**	**−3.5**	**−9.8**
gold boxes rescue (ΔR)	AIME	**−17.2**	**−18.1**	**−12.1**	+0.9	+1.7	**−11.2**	**−11.2**
	GPQA	**−9.1**	**−9.5**	+0.1	+1.9	+2.3	+0.5	−3.3
gold boxes, PRECLOSED level	AIME	**+31.0**	**+31.0**	**+24.1**	−3.5	−3.5	**+24.1**	**+24.1**
dip, steps 25→75	AIME	**−6.9**	**+16.4**	**+16.4**	**+28.4**	**+5.2**	**−6.9**	**+16.4**
held-out panel								
pre-closing hurts, step 200	GPQA-main	**−11.6**	**−10.8**	−0.2	0.0	−0.8	−1.0	−3.6
	AIME 2024	**−30.0**	**−30.0**	**−15.8**	−5.8	−5.8	**−15.8**	**−17.5**
the effect grows, 25→200	GPQA-main	**−11.0**	**−28.0**	**−18.0**	**−17.8**	−0.8	−1.0	**−21.0**
	AIME 2024	**−36.7**	**−35.8**	**−21.7**	**−11.7**	**−12.5**	**−22.5**	**−23.3**
better contract, step 25	GPQA-main	−0.6	**+17.2**	**+17.8**	**+17.8**	0.0	0.0	**+17.4**
	AIME 2024	+6.7	**+5.8**	**+5.8**	**+5.8**	+6.7	+6.7	**+5.8**
gpt-oss: early cost	GPQA-main	−2.2	**−5.4**	**−5.2**	**−6.0**	−2.8	−2.0	**−5.8**
	AIME 2024	+2.5	0.0	+0.8	0.0	+2.5	+3.3	+0.8

involves one of them. Simultaneous intervals (max-$|t|$ over the 66 pairs of each reader, critical value 3.20, or over all 132 differences) keep 6 of the 13 (Table 20); the registered analyzer's own bootstrap marks 8 as marginally robust, this recomputation 7. A lenient regex and the judge-free reader reproduce the delivered ranking almost exactly.

Table 20: **Only on GPQA-main under NATIVE does strict grading reorder public models robustly, and a lenient regex reorders none of those pairs.** Inversions: model pairs ordered one way by the reader and the other way by delivered accuracy (anchor excluded). Robust: both differences' CIs exclude zero, marginally or with simultaneous (max-$|t|$) intervals over 66 or 132 differences; τ_b: Kendall's rank correlation with delivered accuracy.

			strict					regex	judge-free (b)
task, contract	models	pairs	inversions	marginal	sim. 66	sim. 132	τ_b	inversions	inversions
GPQA-main, NATIVE	12	66	13	7	6	6	0.585	0	1
GPQA-main, PRECLOSED	9	36	0	0	0	0	0.986	0	0
AIME 2024, NATIVE	12	66	3	0	0	0	0.880	0	0
AIME 2024, PRECLOSED	9	36	2	0	0	0	0.870	2	0

I.3 REVERSALS WITHIN A TRAJECTORY

Of the 58 fine-tuned trajectories with at least two checkpoints in the Qwen3 and R1-Qwen panels, 27 show a significantly positive gap followed by a significantly negative one within the same contract (Table 21), all in Qwen3. None shows the reverse order, and no R1-Qwen trajectory is ever significantly positive. This summary does not pool the separate exploratory Gemma panel. R1-Llama and Qwen3-4B-Base, evaluated after this analysis, are not included; their cell tables show the same order, e.g. R1-Llama's Gemini GPQA-main NATIVE gap goes from $+11.6$ at step 25 to -13.8 at steps 100 and 200, and Qwen3-4B-Base's gpt-oss GPQA-main PRECLOSED gap from $+19.8$ $[+16.2, +23.4]$ to -3.8 $[-5.6, -2.0]$ (Tables 38 and 40).

Table 21: **Within a fixed contract the gap moves from significantly positive to significantly negative in 27 trajectories and never the other way.** Trajectory: one student, recipe, run, task and contract, followed over its evaluated checkpoints (fine-tuned checkpoints only; learning-rate runs, the every-five-steps rerun and the retrains included). Positive: some checkpoint's gap has a 95% CI above zero; $+ \to -$: a checkpoint with a significantly positive gap precedes one with a significantly negative gap.

student	recipe	trajectories	ever positive	$+ \to -$	$- \to +$
Qwen3-30B-A3B (ours)	Gemini	26	22	15	0
	gpt-oss	16	16	8	0
	heuristic box	4	4	3	0
	gold box	4	4	1	0
R1-Distill-Qwen-7B	Gemini	4	0	0	0
	gpt-oss	4	0	0	0
all		58	46	27	0

I.4 The sign rule where both counts are present

Across all 238 cells the larger count gives the sign of 173 of the 187 nonzero gaps; where both counts are at least 2 pp it does so in 35 of 43 (Table 22), and 7 of the 8 misses are in the added students, where both counts are often moderate and nearly equal. Where both counts are nonzero it gives the sign of 99 of 110; where one is zero the identity fixes the sign up to δ (74 of 77 of those cells follow it; the three exceptions, $|\text{gap}| \leq 0.2$ pp, are single disagreement rows). The judge-free bound holds in 237 of the 238 cells; the exception is the held-out P1 cell. Table 2's public and R1-Llama sign counts use the registered eligibility rules (P-B: $|\text{artifact}| > 0.5$ pp; A1: $\max(a, b) \geq 5$ and $|a - b| \geq 5$), not 'nonzero'.

Table 22: **The larger count gives the sign far less often when both counts are present, but then the gap is small.** Cells from Figure 3 (original program, held-out panel, public models, added students). Every one of the 8 misses with both counts at least 2 pp has $|\text{gap}| \leq 2.5$ pp.

cells	n	nonzero gap	larger count gives sign	share
all cells	238	187	173	93%
both counts > 0	117	110	99	90%
both counts ≥ 2 pp	44	43	35	81%
and $\lvert\text{gap}\rvert \geq 2$ pp	32	32	31	97%
and $\lvert\text{gap}\rvert < 2$ pp	12	11	4	36%
both counts ≥ 5 pp	18	18	13	72%
both counts ≥ 2 pp, by setting				
original program	17	16	15	94%
held-out panel	3	3	3	100%
added students	24	24	17	71%
public models	0	0	0	—
all cells, by setting				
original program	112	97	91	94%
held-out panel	20	15	14	93%
public models	42	16	16	100%
R1-Qwen	36	34	29	85%
R1-Llama	28	25	23	92%

I.5 Checkpoint selection

Strict accuracy picks a different checkpoint than delivered accuracy in 23 of 58 trajectories, whereas the judge-free reader changes the choice in only 8 trajectories. Table 23 separates in-sample comparisons from evaluation on held-out decoding seeds.

I.6 Equivalence bounds for the rescue

Strict reports a rescue of $+31.0$ points on the PRECLOSED AIME level (95% CI from 19.0); delivered accuracy changes by -3.5, and its 95% CI stops at $+4.3$, so a delivered gain of 5 points or more is excluded. The ΔR intervals exclude a delivered rescue larger than 7.8 (AIME) and 2.9 points (GPQA) (Table 24). Equivalence within ± 5 points is not shown: the intervals are too wide.

Table 23: **Choosing checkpoints by strict accuracy costs delivered accuracy; the judge-free reader costs far less.** 58 fine-tuned trajectories (as in Table 21). In sample: each reader picks the best checkpoint on the same rows on which delivered accuracy is measured (biased upward, since delivered accuracy picks on the same noise); split seeds (56 trajectories with at least two decoding seeds): pick on half of the seeds, measure delivered accuracy on the other half, both directions averaged (can be negative). Judge-free (b): the pipeline of Appendix H.

	strict	judge-free (b)
picks a different checkpoint than delivered accuracy	23	8
in sample: delivered accuracy lost > 0	21	8
median loss where > 0 (pp)	21.1	2.1
largest loss (pp)	40.5	10.8
loss with 95% CI above 0	9	0
split seeds: mean loss (pp)	7.4	0.6
trajectories with a loss / a gain	24 / 6	7 / 5
loss with 95% CI above 0	10	0
largest loss (pp)	37.1	21.6

Table 24: **The delivered-accuracy intervals rule out the rescue that strict reports.** Gold-box retrain minus the Gemini recipe at step 200 (seed 42). Rescue: a higher PRECLOSED AIME level, or a negative ΔR (contract reliance $R = \text{NATIVE} - \text{PRECLOSED}$ falls). Largest rescue, delivered: the largest rescue inside the delivered 95% CI; smallest rescue, strict: the smallest inside the strict 95% CI. pp; paired item-cluster bootstrap.

	strict	delivered [95% CI]	largest rescue, delivered	smallest rescue, strict	strict − delivered [95% CI]
level, PRECLOSED AIME	$+31.0$	-3.5 $[-12.1, +4.3]$	4.3	19.0	$+34.5$ $[+20.7, +49.1]$
ΔR, AIME	-17.2	$+0.9$ $[-7.8, +11.2]$	7.8	9.5	-18.1 $[-30.2, -7.8]$
ΔR, GPQA	-9.1	$+1.9$ $[-2.9, +6.8]$	2.9	4.2	-11.0 $[-14.0, -8.2]$

I.7 ANSWERS STATED BEFORE THE CUT

Delivered accuracy gives every budget-hit row no answer. A budget-hit row may, however, have closed its thought and stated an answer in its visible channel before it was cut (*B-closed*: every PRECLOSED budget-hit row, and NATIVE rows with a closer), whereas a *B-open* row never left its thought. Four judge-free readings add credit to delivered accuracy for B-closed rows whose visible channel states the correct answer: + *stated answers* (the last box with an integer or letter, else the last "answer is" or "Answer:" phrase), the same with boxes only, the same with the decision table's lenient regex, and + *never-revised answers* (only rows whose stated answers never change). Because under PRECLOSED the reasoning itself runs in the visible channel, the first three credit truncated reasoning on the PRECLOSED side only; the *symmetric control* also credits B-open rows whose open thought states the correct answer. Table 25 gives the decisions under the primary readings (exploratory, post hoc). No decision changes sign. The step-200 cost, the rescue and the dip keep their verdicts; the step-25 comparison and the interaction keep theirs and grow under the one-sided readings, while the symmetric control keeps them near delivered accuracy. Only the gpt-oss early cost can lose its verdict: from the delivered -9.6 pp it shrinks to -5.1 pp when never-revised answers are credited and to -1.5 pp, whose CI covers zero, when every stated answer is; on the held-out panel every such reading's CI covers zero. Table 26 shows why the reading matters there: the truncated PRECLOSED rows mostly restate an answer in a loop until the budget runs out. We do not judge these 2,771 rows: the judge reads only the last 3,000 characters and would return the answer at the cut rather than the first answer stated, so the judge-free readings above are the appropriate instrument.

Table 25: **Crediting answers stated before the cut reverses no decision; only the gpt-oss early cost can lose significance.** Decisions of Table 1 (original items) and the gpt-oss early cost on the held-out panel and at seed 42. Readers as in the text; bold: 95% CI excludes 0; shaded: verdict differs from delivered accuracy's.

decision	task	strict	delivered	+ stated answers	+ never-revised answers	symmetric control
pre-closing hurts, step 200	GPQA	**−14.3**	−0.9	−0.8	−0.8	−0.8
	AIME	**−15.1**	−0.9	−0.9	−0.9	−2.2
the effect grows, 25→200	GPQA	**−14.2**	**−17.6**	**−39.3**	**−28.9**	**−13.5**
	AIME	**−18.1**	**−9.1**	**−31.5**	**−22.0**	**−6.9**
better contract, step 25	GPQA	−0.2	**+16.6**	**+38.5**	**+28.1**	**+12.7**
gpt-oss: early cost	GPQA	**−3.9**	**−9.6**	−1.5	**−5.1**	**−3.2**
gold boxes rescue (ΔR)	AIME	**−17.2**	+0.9	0.0	0.0	+1.7
	GPQA	**−9.1**	+1.9	+1.6	+2.0	+1.5
gold boxes, PRECLOSED level	AIME	**+31.0**	−3.5	−2.6	−2.6	−2.6
dip, steps 25→75	AIME	**−6.9**	**+28.4**	**+28.4**	**+28.4**	**+6.9**
gpt-oss: early cost, held-out	GPQA-main	−2.2	**−6.0**	−1.2	−2.4	−2.4
gpt-oss: early cost, seed 42	GPQA	−1.8	**−9.3**	+0.2	**−6.0**	−1.4

Table 26: **The gpt-oss student's truncated rows at step 25 mostly loop on a stated answer.** GPQA, both fresh training seeds pooled (1,576 rows per contract). Percentages of all rows unless stated.

	PRECLOSED	NATIVE
hit the budget (% of rows)	51.6	37.2
with a visible channel (B-closed)	51.6	31.4
and a stated answer	47.5	29.7
stated answer correct (% of answered)	47.7	49.1
two or more different answers (% of answered)	43.1	43.2
statements of an answer, median	18	26
characters after the last statement, median	3,886	2,412

Part IV Robustness and generality

Registered predictions. P1–P9: held-out items (Appendix J); P-A–P-C: public models (Appendix L); S1–S6: R1-Qwen and R1-Llama (Appendices M.1 and M.2); A1–A3: R1-Llama addendum (Appendix M.2); Q1–Q4 and, for a second training seed, Q5–Q7: Qwen3-4B-Base (Appendix N). The fresh-seed dip prediction (Appendix F.5) was frozen before its outputs existed but after the seed-42 dip was known.

J PRE-REGISTERED HELD-OUT TEST

Every delivered-answer analysis of the AIME-2025 and GPQA-Diamond generations is post hoc: the visible-channel rule, the judge protocol, the bound and the mechanisms were fixed after reading them. We therefore froze all rules and wrote nine falsifiable predictions before generating a single token on items never evaluated in this project: AIME 2024 (30 items, 4 seeds) and the 250 GPQA *main* items that are not in Diamond (2 seeds), under both contracts, for the base model and the Gemini and gpt-oss checkpoints at steps 25 and 200 (6,200 rows). The pre-registration document, its hash, the frozen panel's row identities and hash, the analyzer, its code locks and the predictions are part of the supplementary material (the panel's prompts contain benchmark text and are not redistributed). The protocol and predictions were fixed before generation; the analyzer was finalised while generation was underway and before any output was read, and one amendment (hardening the analyzer's completeness checks, adding a labelled equivalence analysis for P4 and disclosing that P5's threshold is stricter than the original program's own gpt-oss value, 7.9–9.0%) is recorded with its timing (21 base-model shards existed when the amendment was written, 22 when the second code lock was taken a minute later; none had been read); no threshold was changed. The predictions, with thresholds taken from the original bed and loosened where the held-out items are easier except where noted, are: (P1) the identity holds exactly, no cell's artifact exceeds in absolute value the sum of the bound's two judge-free counts (the registration also defined the δ-augmented form, but registered P1 on this judge-free form), and the 2-point rule holds (every cell whose two masses sum to less than $2\,\mathrm{pp}$ has an artifact of at most $2\,\mathrm{pp}$ in absolute value); (P2) at step 25 budget hits are at least 30% on GPQA-main and 70% on AIME-2024 in every cell, and the Gemini NATIVE unclosed share is at least 10%; (P3) strict$-$committed is at least $+5\,\mathrm{pp}$ (CI excluding 0) in every step-25 GPQA cell and $+10\,\mathrm{pp}$ on AIME; (P4) for Gemini at step 200 under PRECLOSED the unparsed-commitment share is at least 10% on GPQA and 50% on AIME, the strict contract contrast on GPQA is at most $-5\,\mathrm{pp}$ with CI excluding 0 while the committed contrast lies within $\pm 5\,\mathrm{pp}$ with CI including 0; (P5) gpt-oss at step 200 has unparsed share at most 5% and $|\mathrm{strict} - \mathrm{committed}| \leq 3\,\mathrm{pp}$; (P6) the base has $|\mathrm{strict} - \mathrm{committed}| \leq 1.5\,\mathrm{pp}$ and unclosed share at most 1%; (P7) the strict recipe gap at step 200 under PRECLOSED on GPQA exceeds the committed gap by at least $5\,\mathrm{pp}$; (P8) correctness given a box is ordered closed $>$ unclosed $>$ budget at step 25, each gap at least $5\,\mathrm{pp}$; (P9) the Gemini student's hedging rate exceeds the gpt-oss student's by at least $15\,\mathrm{pp}$ at step 200 (unassessable, not failed, if either recipe has fewer than 30 contested rows).

J.1 OUTCOME

The locked analyzer was run once, after all 6,200 rows and both judge files were complete. Six of the nine predictions pass as registered (P3, P4, P6, P7, P8, P9) and three fail (P1, P2, P5). Table 27 gives every cell; Table 28 gives every prediction with its constituent quantities. No threshold was moved. Amendment 1 named three failure classes (measurement failure, transfer failure, insufficient support); we use them where they fit and say plainly where they do not.

P1: fail. The identity equation 1 holds to floating-point precision in all 20 cells. The judge-free form of the bound—the two masses a and b without the δ term, the form on which we chose to register P1 although the same document defined the δ-augmented form—is violated in one cell by one row: base, GPQA-main, PRECLOSED, where the response labelled the four quantum states of the question (a)–(d), concluded "the only entangled states among the given options are (a) and (c)" and wrote "Answer: `\boxed{B}`", which is correct; the judge returned A. Both judge-free counts are zero in that cell and the artifact is $+0.2\,\mathrm{pp}$, so $\delta = 0.2\%$ there (one of the cell's 500 rows). By the registration's own falsification clause this falsifies the bound *as stated there*. The form the paper states, equation 3, bounds each side by one of the two masses plus δ and was adopted

Table 27: Held-out panel, every cell (%; $n = 120$ per AIME-2024 cell, 500 per GPQA-main cell; base: the untouched model, which has no step; est.: estimate; brackets: unadjusted item-cluster 95% CI). B/U/C: termination states (hit the budget, ended inside an unclosed thought, closed). $a = P(\pi, \mathrm{B} \cup \mathrm{U})$ and $b = P(\mathrm{C}, \neg\pi)$ are the two judge-free counts of equation 3; $\delta = P(\mathrm{C}, \pi, y_\pi \neq y_c)$ is recomputed from the released rows per equation 3 (the locked analyzer's disagreement field compared correctness indicators on committed rows and understated three cells by 0.2–0.6 pp; both versions are in the supplement, and the bound holds with either, with the recomputed δ by construction); unparsed commitment is the share of rows that commit to an answer the strict rule does not parse. CIs: item-cluster bootstrap on the paired artifact. π: a strict parse is present; y_π, y_c: the parsed and the delivered answer; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; artifact: strict minus committed; held-out panel: AIME 2024 and 250 GPQA-main items never evaluated before; step: optimizer step; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model); pp: percentage points.

			state					artifact				unparsed
recipe	step	contract	B	U	C	strict	committed	est. [95% CI]	a	b	δ	commit.
AIME-2024												
base		NATIVE	5.0	0.0	95.0	90.8	90.8	0.0 [0.0, 0.0]	0.0	0.0	0.0	0.0
base		PRECLOSED	8.3	0.0	91.7	85.8	85.8	0.0 [0.0, 0.0]	0.0	0.0	0.0	0.0
Gemini	25	NATIVE	87.5	8.3	4.2	22.5	3.3	+19.2 [+10.0, +29.2]	38.3	0.0	0.0	0.0
Gemini	25	PRECLOSED	89.2	0.0	10.8	29.2	9.2	+20.0 [+10.0, +30.8]	40.0	0.8	0.0	0.0
Gemini	200	NATIVE	29.2	0.0	70.8	40.8	58.3	−17.5 [−26.7, −9.2]	0.0	24.2	0.0	24.2
Gemini	200	PRECLOSED	35.0	0.0	65.0	10.8	52.5	−41.7 [−55.8, −28.3]	0.0	52.5	0.0	51.7
gpt-oss	25	NATIVE	65.0	1.7	33.3	36.7	24.2	+12.5 [+5.8, +19.2]	26.7	5.0	0.0	2.5
gpt-oss	25	PRECLOSED	70.0	0.0	30.0	39.2	24.2	+15.0 [+8.3, +22.5]	31.7	5.8	0.0	2.5
gpt-oss	200	NATIVE	17.5	0.0	82.5	79.2	79.2	0.0 [0.0, 0.0]	0.0	0.0	0.0	0.0
gpt-oss	200	PRECLOSED	14.2	0.0	85.8	80.0	80.0	0.0 [−2.5, +2.5]	0.8	1.7	0.0	0.8
GPQA-main												
base		NATIVE	0.0	0.0	100.0	65.8	65.8	0.0 [0.0, 0.0]	0.0	0.0	0.0	0.0
base		PRECLOSED	0.0	0.0	100.0	66.0	65.8	+0.2 [0.0, +0.6]	0.0	0.0	0.2	0.0
Gemini	25	NATIVE	37.0	27.2	35.8	53.4	23.0	+30.4 [+26.0, +34.8]	54.2	1.4	0.2	0.4
Gemini	25	PRECLOSED	34.6	0.0	65.4	52.8	40.8	+12.0 [+9.0, +15.0]	28.6	4.6	0.4	3.4
Gemini	200	NATIVE	0.8	1.8	97.4	56.6	56.2	+0.4 [−0.6, +1.4]	1.6	1.0	0.0	0.8
Gemini	200	PRECLOSED	0.4	0.0	99.6	45.0	56.2	−11.2 [−14.2, −8.4]	0.2	18.8	0.0	18.8
gpt-oss	25	NATIVE	32.8	0.6	66.6	53.6	40.6	+13.0 [+10.0, +16.4]	27.0	1.0	0.4	0.8
gpt-oss	25	PRECLOSED	42.4	0.0	57.6	51.4	34.6	+16.8 [+13.4, +20.4]	35.4	1.8	0.8	1.0
gpt-oss	200	NATIVE	9.2	0.0	90.8	56.6	61.8	−5.2 [−7.4, −3.2]	0.6	8.0	0.0	8.0
gpt-oss	200	PRECLOSED	7.8	0.0	92.2	59.8	62.2	−2.4 [−4.0, −0.8]	0.6	5.8	0.4	5.6

Table 28: Held-out predictions as registered, observed values and outcome. Failures are classified in the text; no threshold was moved. P3, P4 and P7, the predictions that test whether the artifacts transfer, pass; P1 fails by one row, and P2 and P5 fail their registered quantitative thresholds. P1–P9: the nine predictions registered before generation (text above); identity: equation 1; judge-free form: $|\text{artifact}| \leq a+b$ with the judge-free counts a (rows with a strict parse that hit the budget or ended unclosed) and b (closed rows without one); δ: parser–judge disagreement mass, by which the δ-form widens the bound; 2-point rule: every cell with $a+b < 2\,\text{pp}$ has $|\text{artifact}| \leq 2\,\text{pp}$; artifact: strict minus committed; strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; unparsed commitment: delivered answers the strict rule does not parse; correct-given-boxed: accuracy of rows with a strict parse, by termination state; hedging: contested tails naming two or more option letters; immune: no strict–committed artifact; pooled: both recipes together; robust boxed: the last balanced box with `\text{}` wrappers stripped; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model); pp: percentage points; CI: item-cluster bootstrap 95% interval. δ-form: the bound of equation 3 with the disagreement mass δ added; identity error: the largest absolute deviation from equation 1.

	registered	observed	outcome
P1	identity exact; judge-free form $\|\texttt{artifact}\| \leq a+b$ never violated; 2-point rule	identity error $< 10^{-14}$ in 20/20; judge-free form violated in 1 cell by 1 row (base GPQA-main PRECLOSED, $+0.2$ pp, $\delta = 0.2\%$: one parser–judge disagreement); δ-form holds by construction; 2-point rule 5/5	FAIL (falsifies the bound as registered)
P2	step-25 budget $\geq 30\%$ GPQA-main, $\geq 70\%$ AIME-2024, all four cells; Gemini NATIVE unclosed $\geq 10\%$	GPQA-main 37.0/34.6/32.8/42.4; AIME-2024 87.5/89.2/65.0/70.0 (one exactly at threshold); unclosed 27.2	FAIL (one cell, $65.0 < 70$)
P3	strict$-$committed $\geq +5$ pp (CI > 0) in every step-25 GPQA-main cell, $\geq +10$ pp AIME-2024	GPQA-main $+30.4/+12.0/+13.0/+16.8$, all CIs > 0; AIME-2024 $+19.2/+20.0/+12.5/+15.0$	PASS
P4	Gemini, step 200, PRECLOSED unparsed commitment $\geq 10\%$ GPQA-main, $\geq 50\%$ AIME-2024; GPQA-main strict contrast ≤ -5 (CI < 0), committed within ± 5 (CI $\ni 0$)	18.8%, 51.7% (62 of 120 rows vs. 60); strict -11.6 $[-16.4, -6.8]$, committed 0.0 $[-4.4, +4.4]$	PASS
P5	gpt-oss, step 200, unparsed commitment $\leq 5\%$ and $\|\texttt{artifact}\| \leq 3$ pp in all four cells	AIME-2024 0.0/0.8%, 0.0/0.0 pp; GPQA-main 8.0/5.6%, $-5.2/-2.4$ pp (robust boxed, exploratory: 1.6/0.4%)	FAIL (gpt-oss not immune on GPQA-main)
P6	base $\|\texttt{artifact}\| \leq 1.5$ pp, unclosed $\leq 1\%$	0.0/0.0/0.0/$+0.2$ pp; unclosed 0	PASS
P7	strict (gpt-oss$-$Gemini) exceeds committed by ≥ 5 pp at step 200, PRECLOSED, GPQA-main	14.8 vs 6.0 pp	PASS
P8	correct-given-boxed closed $>$ unclosed $>$ budget, gaps ≥ 5 pp (pooled)	$63.0 > 57.9 > 52.6$ (gaps 5.1, 5.2); within Gemini unclosed$-$budget $= 0.6$	PASS (pooled; middle step not within-recipe)
P9	Gemini hedging exceeds gpt-oss by ≥ 15 pp (≥ 30 rows each)	39.8 ($n = 98$) vs 7.4% ($n = 68$)	PASS

after the independent review and before any held-out row was read; it follows from the definitions (Appendix H), so it holds in every cell by construction and is not tested here. The larger judge-free count gives the sign of 14 of the 15 nonzero artifacts; the fifteenth is this cell, where the two masses are equal. The 2-point rule holds (five cells with $a + b < 2$, largest artifact 0.2 pp). We score P1 as registered.

P2: fail (a wrong quantitative prediction in one of nine quantities; none of the three named classes fits). Budget-hit shares at step 25 are 37.0, 34.6, 32.8 and 42.4% on GPQA-main (registered ≥ 30) and 87.5, 89.2, 65.0 and 70.0% on AIME-2024 (registered ≥ 70; the base hits the budget on 5.0 and 8.3%), and the Gemini student's NATIVE unclosed share on GPQA-main is 27.2% (registered ≥ 10). The gpt-oss NATIVE AIME cell misses by 5 pp on 120 rows, and the gpt-oss PRECLOSED AIME cell passes exactly at the threshold. The collapse is present in every cell; its size in one cell falls short of the registered threshold. (Under greedy decoding the same cell hits the budget on 96.7% of rows, mostly in verbatim loops; exploratory, below, and without bearing on this verdict.)

P3: pass. Strict minus committed at step 25 is $+30.4$ (CI $[+26.0, +34.8]$), $+12.0$ $[+9.0, +15.0]$, $+13.0$ $[+10.0, +16.4]$ and $+16.8$ pp $[+13.4, +20.4]$ on GPQA-main (Gemini NATIVE, Gemini PRECLOSED, gpt-oss NATIVE, gpt-oss PRECLOSED) and $+19.2$, $+20.0$, $+12.5$ and $+15.0$ pp on AIME-2024, with each interval excluding zero on both evaluation tasks.

P4: pass. For Gemini at step 200 under PRECLOSED the committed-but-unparsed share is 18.8% on GPQA-main and 51.7% on AIME-2024; the GPQA-main contract contrast is -11.6 pp strict (CI $[-16.4, -6.8]$) and 0.0 pp committed ($[-4.4, +4.4]$). The labelled equivalence analysis of Amendment 1 (committed CI inside $[-5, +5]$) also passes. The AIME share clears its threshold by just two rows: 62 of the 120 responses, against the registered minimum of 60.

P5: fail (the gpt-oss recipe is not grader-immune on GPQA-main under the strict rule; the unparsed-share half of the threshold was disclosed as demanding in Amendment 1, the artifact half is a substantive miss). On AIME-2024 the gpt-oss step-200 cells have unparsed commitment 0.0 and 0.8% and artifacts of 0.0 pp. On GPQA-main the unparsed share is 8.0% (NATIVE) and 5.6% (PRECLOSED) against a registered ≤ 5, and the artifact is -5.2 pp ($[-7.4, -3.2]$) and -2.4 pp ($[-4.0, -0.8]$) against $|\cdot| \leq 3$. Of the 68 unparsed commitments on GPQA-main, 54 wrap the letter as "`\boxed{\text{X}}`", four as "`\boxed{\,X\,}`", seven box the letter together with a value, a formula or the option's wording, and three are neither (an option E, a box without braces, an expression); under the labelled robust-boxed reading the unparsed share is 1.6 and 0.4%. Both halves of the threshold were stricter than the original program's own gpt-oss step-200 GPQA values (unparsed commitment 7.9–9.0%, artifact -5.2 and -6.5 pp); Amendment 1 disclosed the first before any output was read, and the registered immunity prediction fails again on GPQA-main. The substantive reading is that the gpt-oss recipe shows no artifact on AIME, where it boxes integers, but does on GPQA-main, where a nested-box habit costs it 2–5 strict points: a further instance of the confound.

P6: pass. Base artifacts are 0.0, 0.0, 0.0 and $+0.2$ pp; the unclosed share is 0 in all four cells.

P7: pass. At step 200 under PRECLOSED on GPQA-main the gpt-oss$-$Gemini gap is 14.8 pp strict and 6.0 pp committed, an excess of 8.8 pp that clears the pre-registered minimum excess of 5 pp.

P8: pass, at the threshold. Pooled over recipes at step 25 (NATIVE, GPQA-main), correctness given a box is 63.0% in closed rows ($n = 500$), 57.9% in unclosed rows ($n = 121$) and 52.6% in budget-hit rows ($n = 285$): gaps of 5.1 and 5.2 pp. Per recipe (Amendment 1): Gemini 66.3 / 56.8 / 56.2% ($n = 172$, 118, 153); gpt-oss 61.3 / 100 / 48.5% ($n = 328$, 3, 132). The pooled ordering is as registered, but its middle step is a composition effect: within Gemini the unclosed–budget gap is 0.6 pp, and gpt-oss contributes 132 of the 285 budget-hit rows at 48.5% against three unclosed rows, so unclosed $>$ budget is not supported within either recipe. Closed $>$ budget holds in both recipes; the gpt-oss closed–unclosed comparison is inconclusive with three parsed unclosed rows.

P9: pass. Among contested GPQA-main rows at step 200 (as registered, both contracts pooled; the analyzer counts closed rows without a strict parse on which the judge commits to an answer), tails naming two or more option letters are 39.8% for Gemini ($n = 98$) and 7.4% for gpt-oss ($n = 68$).

J.2 WHAT TRANSFERRED

Beyond the registered tests, the panel repeats the paper's headline readings on items never evaluated before. The Gemini step-200 contract effect is -11.6 pp strict and 0.0 pp committed on GPQA-main, and -30.0 ($[-43.3, -17.5]$) strict against -5.8 pp ($[-12.5, +0.8]$) committed on AIME-2024. The step-25 reattribution recurs: Gemini on GPQA-main is -0.6 pp strict and $+17.8$ pp ($[+13.2, +22.4]$) committed, with 27.2% of NATIVE rows ending inside an unclosed thought. The gpt-oss recipe's early cost of pre-closing recurs under the registered visible-channel reading (budget-hit and unclosed rows commit nothing): -6.0 pp committed ($[-11.0, -1.0]$) against -2.2 pp strict ($[-6.6, +2.2]$); crediting answers stated in truncated rows, every CI covers zero (Appendix I.7). The step-200 recipe gap under PRECLOSED is 14.8 strict against 6.0 pp committed on GPQA-main and 69.2 against 27.5 pp on AIME-2024. Along base→step 25→step 200 the Gemini NATIVE AIME-2024 curve reads $90.8 \rightarrow 22.5 \rightarrow 40.8\%$ strict and $90.8 \rightarrow 3.3 \rightarrow 58.3\%$ committed. These readings are descriptive; the registered outcomes are the nine above.

J.3 SECOND JUDGE FAMILY (SENSITIVITY, NOT REGISTERED)

Re-judging the 4,870 terminated held-out rows with gemma-3-12b-it under the same prompt, the two judge families agree on 99.71% of rows and on 98.44% of the 320 closed rows without a strict parse; no cell's delivered accuracy and no contract contrast moves by more than 0.83 pp, and the point values of P3, P4, P6 and P7 pass and P5 fails under either judge. On the one P1-violating row the second judge returns B, agreeing with the parser and with gold, so under the second judge the bound as registered would hold in all 20 cells; the registered outcome uses the registered judge and stays a failure.

J.4 GREEDY DECODING (EXPLORATORY, NOT REGISTERED)

The protocol samples at $T{=}1$, and an early checkpoint might fail to stop only because sampling is noisy. We decoded the held-out prompts greedily, every item once under both contracts (60 AIME-2024 and 500 GPQA-main rows per checkpoint), for the base model and both step-25 checkpoints (Table 30), with the budget, engine version, GPU type, tensor parallelism, model configurations and frozen inputs of the registered runs but up to 128 rather than 64 concurrent sequences. Removing the sampling noise does not relieve the collapse: budget hits are 80.4–100.0% for Gemini (sampled: 34.6–89.2) and 90.4–96.7% for gpt-oss (32.8–70.0). The base model does not collapse: it hits the budget on 2 of 560 rows and scores 93.3 and 90.0% on AIME-2024 and 65.6 and 63.2% on GPQA-main. The failure changes form. Sampled budget-hit rows are drifting derivations (§6); under greedy decoding at step 25, 63–100% of budget-hit rows are verbatim loops (fewer than half of the 8-grams in the last 4,096 tokens are distinct), and no greedy row stops inside an unclosed thought. Verbatim repetition is a known failure of greedy decoding (Holtzman et al., 2020), and the base model's card recommends $T{=}0.6$, top-p 0.95 and top-k 20 and notes that a presence penalty can reduce endless repetitions; but the base model decoded greedily loops on 1 of its 560 rows, so the loops at step 25 are not produced by greedy decoding alone, in line with reports that distilled students loop where their teachers rarely do (Pipis et al., 2026). At that setting ($T{=}0.6$, top-p 0.95, top-k 20; one sampled row per item and contract, 64 concurrent sequences) the step-25 checkpoints hit the budget on 46.8–100.0% (Gemini) and 77.6–90.0% (gpt-oss) of the rows of a cell, 60–100% of those rows loop and none stops inside an unclosed thought, while the base model hits the budget on 5 of its 560 rows; in each of the eight step-25 cells the budget-hit share rises, or stays at 100%, as the temperature falls from 1 to 0.6 to 0. At $T{=}0.6$, 0–3% (Gemini) and 26–39% (gpt-oss) of a cell's budget-hit rows hold a parsable box. Greedy budget-hit rows also carry little truncation credit: none of the Gemini student's 471 budget-hit rows contains a parsable box, while 3–13% of the gpt-oss student's do, against 40–83% of the $T{=}1$ step-25 budget-hit rows on the same items. So 25 SFT steps impair stopping under all three decodings while the base model rarely fails to stop under any; decoding changes the form and the rate of the failure, not its presence. The greedy and model-card rows were later scored with the registered judge (Table 31; exploratory, post hoc): at step 25 the gpt-oss student keeps a positive gap under every decoding, while the Gemini student's nearly vanishes. No registered verdict depends on these rows.

J.5 PRESENCE PENALTY AND A DOUBLED BUDGET (EXPLORATORY, NOT REGISTERED)

Adding the presence penalty the card suggests (1.5) to its recommended sampling lowers the step-25 budget-hit share in every cell but leaves it at 22.8–76.7% (Gemini) and 26.4–80.0% (gpt-oss), against 3 of 560 rows for the base model; 77–99% of the GPQA budget-hit rows still loop verbatim (35–53% on AIME), and no row stops inside an unclosed thought; 0–4% (Gemini) and 26–71% (gpt-oss) of a cell's budget-hit rows hold a parsable box. Doubling the budget to 65,536 tokens at the registered $T{=}1$ (first planned seed; 30 AIME-2024 and 250 GPQA-main rows per contract and checkpoint) leaves 63–70% of step-25 AIME rows at the budget: of the rows that hit 32,768 tokens in the registered runs, 26–39% (AIME) and 57–79% (GPQA) end by 65,536, and rows that still hit it rarely loop (0–14%). The 64k rows are fresh samples (none continues its registered row verbatim), so these are rates, not per-row trajectories. A larger budget recovers part of GPQA but not AIME: the step-25 AIME collapse is not an artifact of the 32,768-token budget.

J.6 MODEL-CARD DECODING AT STEP 200 (EXPLORATORY, NOT REGISTERED)

We also decoded both step-200 checkpoints with the model card's sampling ($T{=}0.6$, top-p 0.95, top-k 20; the first planned seed: 30 AIME-2024 and 250 GPQA-main rows per contract; 32,768-token budget) and scored them with the registered judge, next to the registered $T{=}1$ rows of the same seed (Table 29). The Gemini student still loops: 35.6–76.7% of its rows hit the budget, against 0.8–36.7% at $T{=}1$, and 99–100% of its GPQA budget-hit rows are verbatim loops (55–74% on AIME); one row ends unclosed. The strict grader misreads the contract. On GPQA-main it shows no effect (+0.4 pp) while delivered scoring puts PRECLOSED ahead by 15.6 pp, because the NATIVE arm loops more (artifact −15.2 pp, against −10.8 pp at $T{=}1$); on AIME-2024 strict puts PRECLOSED 13.3 pp behind and delivered scoring 33.3 pp ahead (artifact −46.7 pp). The artifact has two parts: the PRECLOSED arm's unread prose persists (−7.2 pp on GPQA-main and −50.0 pp on AIME-2024, against −10.4 and −43.3 pp at $T{=}1$), and the looping NATIVE arm regains truncation credit on GPQA-main (+8.0 pp, all of its artifact). The gpt-oss student shows small artifacts (−6.7 and −2.0 pp, neither CI excluding zero; budget hits 5.2–23.3%). Termination recovery by step 200 thus holds at $T{=}1$ only, and under the card's decoding the step-200 contract artifact mixes both clocks.

Table 29: Model-card decoding at step 200 (exploratory, not registered): the first planned seed of the held-out panel (AIME-2024: 30 items; GPQA-main: the 250 GPQA items outside Diamond) decoded with the base model card's sampling ($T{=}0.6$, top-p 0.95, top-k 20; card) and, on the same rows, the registered $T{=}1$ decoding. Budget: share of rows that hit the 32,768-token budget (%, NATIVE / PRECLOSED); loop: share of those rows whose last 4,096 tokens have fewer than 50% distinct 8-grams (%, card only; –: not computed). Strict, committed, artifact: PRECLOSED−NATIVE contrasts (pp) of strict accuracy (last `\boxed{}`), delivered accuracy (judge-read visible answer) and strict minus committed, with the artifact's item-cluster 95% CI; bold: CI excludes 0. NATIVE/PRECLOSED: generation starts inside / after the thought; Gemini, gpt-oss: distillation recipes named for their teachers; pp: percentage points.

recipe	decoding	task	budget	loop	strict	committed	artifact	95% CI
Gemini	card	AIME-2024	76.7 / 36.7	74 / 55	**−13.3**	**+33.3**	**−46.7**	[−66.7, −26.7]
Gemini	card	GPQA-main	58.4 / 35.6	99 / 100	+0.4	**+15.6**	**−15.2**	[−20.4, −10.4]
Gemini	$T{=}1$	AIME-2024	36.7 / 23.3	–	−6.7	+6.7	−13.3	[−33.3, +6.7]
Gemini	$T{=}1$	GPQA-main	0.8 / 0.8	–	**−11.6**	−0.8	**−10.8**	[−15.2, −6.8]
gpt-oss	card	AIME-2024	23.3 / 23.3	14 / 29	−3.3	+3.3	−6.7	[−16.7, 0.0]
gpt-oss	card	GPQA-main	9.2 / 5.2	39 / 31	−5.2	−3.2	−2.0	[−4.4, +0.4]
gpt-oss	$T{=}1$	AIME-2024	23.3 / 13.3	–	**+16.7**	**+13.3**	+3.3	[0.0, +10.0]
gpt-oss	$T{=}1$	GPQA-main	8.4 / 7.6	–	−0.8	−2.8	+2.0	[−1.6, +5.6]

J.7 TRAINING OVERLAP

"Held-out" above means never evaluated in this project. Running the frozen lexical contamination detector of Appendix B over both 100,000-question training corpora against the 280 held-out questions found no exact match on GPQA-main or AIME-2024 and one near-match: AIME-2024 item 25 (a triangle inscribed in a circle with $AB = 5$, $BC = 9$, $AC = 10$) appears, reworded, as one training question present in both corpora, with the point labels swapped and asking for the length AD rather than $m + n$. We count it as a contamination of one of the 30 AIME-2024 items for both

Table 30: Greedy decoding against the registered $T{=}1$ cells (%). Budget and unclosed: termination states. Loop: share of greedy budget-hit rows whose last 4,096 tokens have fewer than 50% distinct 8-grams. Box: share of greedy non-answer rows containing a box the strict rule parses. Strict: strict accuracy. Greedy cells have one row per item ($n = 30$ or 250), $T{=}1$ cells four or two ($n = 120$ or 500); –: no budget-hit or non-answer rows; exploratory. $T{=}1$: the registered sampling at temperature 1; greedy: temperature 0; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; Gemini, gpt-oss: distillation recipes named for their teachers (a Gemini-family and a gpt-oss model); step: optimizer step. AIME-2024, GPQA-main: the held-out items (AIME 2024; the 250 GPQA items outside Diamond).

				budget		unclosed		loop	box	strict	
recipe	step	task	contract	$T{=}1$	greedy	$T{=}1$	greedy	greedy	greedy	$T{=}1$	greedy
base		AIME-2024	NATIVE	5.0	3.3	0.0	0.0	0.0	0.0	90.8	93.3
base		AIME-2024	PRECLOSED	8.3	0.0	0.0	0.0	–	–	85.8	90.0
base		GPQA-main	NATIVE	0.0	0.0	0.0	0.0	–	–	65.8	65.6
base		GPQA-main	PRECLOSED	0.0	0.4	0.0	0.0	100.0	0.0	66.0	63.2
Gemini	25	AIME-2024	NATIVE	87.5	100.0	8.3	0.0	63.3	0.0	22.5	0.0
Gemini	25	AIME-2024	PRECLOSED	89.2	100.0	0.0	0.0	80.0	0.0	29.2	0.0
Gemini	25	GPQA-main	NATIVE	37.0	80.4	27.2	0.0	100.0	0.0	53.4	15.2
Gemini	25	GPQA-main	PRECLOSED	34.6	84.0	0.0	0.0	100.0	0.0	52.8	11.6
gpt-oss	25	AIME-2024	NATIVE	65.0	96.7	1.7	0.0	89.7	6.9	36.7	10.0
gpt-oss	25	AIME-2024	PRECLOSED	70.0	96.7	0.0	0.0	96.6	3.4	39.2	6.7
gpt-oss	25	GPQA-main	NATIVE	32.8	90.4	0.6	0.0	100.0	11.1	53.6	14.0
gpt-oss	25	GPQA-main	PRECLOSED	42.4	95.6	0.0	0.0	100.0	13.0	51.4	12.8

Table 31: Step-25 decoding probes on the held-out panel, scored with the registered judge (exploratory, post hoc; one decoding per item: 30 AIME-2024 and 250 GPQA-main rows per cell; %, pp). Greedy: temperature 0; model card: $T{=}0.6$, top-p 0.95, top-k 20; + presence penalty: the same with presence penalty 1.5. Gap: strict − committed.

			NATIVE			PRECLOSED		
recipe	decoding	task	strict	committed	gap	strict	committed	gap
Gemini	greedy	AIME-2024	0.0	0.0	0.0	0.0	0.0	0.0
Gemini	greedy	GPQA-main	15.2	15.2	0.0	11.6	11.6	0.0
Gemini	model card	AIME-2024	6.7	3.3	+3.3	0.0	0.0	0.0
Gemini	model card	GPQA-main	38.8	38.8	0.0	34.8	34.8	0.0
Gemini	card + presence penalty	AIME-2024	23.3	23.3	0.0	36.7	40.0	−3.3
Gemini	card + presence penalty	GPQA-main	54.8	54.8	0.0	47.6	47.2	+0.4
gpt-oss	greedy	AIME-2024	10.0	3.3	+6.7	6.7	3.3	+3.3
gpt-oss	greedy	GPQA-main	14.0	8.0	+6.0	12.8	4.0	+8.8
gpt-oss	model card	AIME-2024	30.0	6.7	+23.3	30.0	13.3	+16.7
gpt-oss	model card	GPQA-main	32.4	17.2	+15.2	32.4	10.0	+22.4
gpt-oss	card + presence penalty	AIME-2024	50.0	33.3	+16.7	53.3	20.0	+33.3
gpt-oss	card + presence penalty	GPQA-main	57.6	50.0	+7.6	54.4	40.0	+14.4

recipes. Removing that item (29 items, 116 rows per AIME cell) changes no registered verdict: the P3 AIME artifacts are $+19.8$, $+20.7$, $+12.9$ and $+14.7$ pp, the P4 AIME unparsed-commitment share is 53.4%, the P6 and P5 AIME cells still pass their halves, and the gpt-oss PRECLOSED AIME budget share, which passed P2 exactly at 70.0%, becomes 69.0% in a prediction that already failed. The Gemini step-200 AIME contract effect is -31.0 pp strict against -6.0 pp committed and the recipe gap 68.1 against 25.0 pp without the item.

J.8 SCORER INVARIANCE UNDER ANSWER-PRESERVING REWRITES (ORIGINAL PROGRAM, NOT REGISTERED)

A second instrument test holds the answer fixed and changes only its presentation. On the original program (AIME-2025 and GPQA-Diamond, not the held-out items), from the seed-42 Gemini and gpt-oss checkpoints at steps 25 and 200 and the base anchor we took every terminated row with a visible answer channel whose commitment is machine-verified (strict parse present and the judge agrees: 6,165 rows; or no parse, both judge families agree and the answer occurs literally in the tail: 396 rows) and rewrote its answer segment so that the verified answer X is stated in five forms (four on GPQA) plus a control: the original; existing boxes unwrapped and a final line "Answer: `\boxed{X}`"; the same with a plain "Answer: X"; with a nested "`\boxed{\text{X}}`"; on

AIME with a boxed equation "\boxed{n = X}"; and, as a control, the answer segment preceded by "Initially I thought the answer was \boxed{Y}, but that is wrong" with a wrong Y (where the original has no strict parse, a plain final line "Answer: X" is appended so that the commitment stays explicit). Table 32 gives the share of rows on which each scorer returns X. The strict rule returns X on every row only under the canonical box; on the distractor control it returns the wrong Y on exactly the rows whose original ending had no parsable box (5.17% of GPQA and 15.71% of AIME rows), because the distractor is then the only box the strict rule can parse. The robust-boxed rule repairs nesting only. The committed judge, run on the same 33,346 variant rows, returns X on 99.97–100% of rows in every variant and returns the distractor Y on 0.02% of GPQA control rows and none on AIME; per source row, it gives the identical correct verdict on all variants for 99.97% of the 6,561 rows, against 0% for the strict and robust-boxed rules. Three caveats: the selection used the judge's own reading of the original row (so the original column is 100% by construction, and the test is of invariance to presentation, not of correctness); the rewrites are deterministic templates, not the full variety of model prose, and change the length of the text the judge reads by at most 81 characters, so they test format, not verbosity (a length panel is in Appendix E); and the distractor sits at the start of the answer segment, which the strict and robust-boxed rules read whole but the judge only in its last 3,000 characters, so the boxed distractor lay inside the judge's window on 4,064 of the 6,020 GPQA and 415 of the 541 AIME control rows. On those rows the judge returns Y once (GPQA), and on the 205 of them where the strict rule returns Y, all of which end in the appended plain line, it returns X every time.

Table 32: Original program (AIME-2025, GPQA-Diamond; not registered): share of rows (%) on which each scorer returns the verified answer X under answer-preserving rewrites; for the distractor control, the share returning the wrong Y in parentheses; robust: the robust-boxed rule, the last balanced box with \text{} wrappers stripped. The boxed-equation rewrite exists for AIME only (n/a on GPQA). Variants: canonical box, the answer as a final line "Answer: \boxed{X}" with earlier boxes unwrapped; plain final line, "Answer: X"; nested box, \boxed{\text{X}}; boxed equation, \boxed{n = X}; earlier boxed distractor, a wrong boxed Y at the start of the answer segment; strict: the frozen rule reading the last \boxed{} answer after the last thought closer (else the whole output); judge: the delivered-answer judge, gold-blind, reading the last 3,000 characters of the visible channel.

	GPQA (n = 6,020)			AIME (n = 541)		
variant	strict	robust	judge	strict	robust	judge
original	94.8	97.1	100.00	84.3	84.3	100.00
canonical box	100.0	100.0	100.00	100.0	100.0	100.00
plain final line	0.0	0.0	99.98	0.0	0.0	100.00
nested box	0.0	100.0	100.00	0.0	100.0	100.00
boxed equation	n/a	n/a	n/a	0.0	0.0	100.00
earlier boxed distractor	94.8 (5.2)	97.1 (2.8)	99.97 (0.02)	84.3 (15.7)	84.3 (14.8)	100.00 (0.00)

K A RERUN WITH CHECKPOINTS EVERY FIVE STEPS

To see where the step-25 collapse begins, the Gemini trajectory was rerun with the same data order, seeds and learning-rate schedule, saving every 5 steps and stopping at step 50 (its learning rate equals the original's at every logged step, and the loss differs by at most 0.005). The ten checkpoints were decoded on AIME-2024 (held-out panel, first planned seed, both contracts, T=1, 32,768 tokens) and scored with the registered judge (Table 33). Steps 5 to 15 are indistinguishable from the base model: at most 6.7% of rows hit the budget and strict equals delivered accuracy (86.7–93.3%) in every cell, as for the base model on the same items (5.0 and 8.3% budget hits, 90.8 and 85.8%). At step 20, 90.0% of NATIVE rows hit the budget, and the strict grader reports 40.0% against 3.3% committed; from step 20 to 35 the NATIVE artifact is +30.0 to +36.7 points, credit for boxes in unfinished thoughts. By step 50 the PRECLOSED artifact has turned negative (−20.0 points) as prose answers go unread. The collapse is therefore a transition within five steps, inside the warm-up, and step 25, the first checkpoint of the original run, falls just after this transition, still inside the warm-up; one decoding per item, so each cell rests on 30 rows.

Table 33: Every-5-steps rerun of the Gemini trajectory on AIME-2024 (exploratory; 30 items, one decoding each, T=1; %). B: budget hit; U: ended inside an unclosed thought; strict: last-`\boxed{}` rule; comm.: delivered accuracy (judge-read visible answer); NATIVE/PRECLOSED: generation starts inside / after the thought; step: optimizer step.

	NATIVE				PRECLOSED		
step	B	U	strict	comm.	B	strict	comm.
5	3.3	0.0	93.3	93.3	6.7	86.7	86.7
10	6.7	0.0	90.0	90.0	0.0	86.7	86.7
15	6.7	0.0	90.0	90.0	0.0	93.3	93.3
20	90.0	6.7	40.0	3.3	50.0	46.7	26.7
25	83.3	10.0	40.0	6.7	76.7	36.7	16.7
30	93.3	6.7	36.7	0.0	73.3	30.0	23.3
35	80.0	20.0	30.0	0.0	70.0	20.0	23.3
40	70.0	20.0	20.0	6.7	50.0	36.7	33.3
45	66.7	0.0	30.0	33.3	73.3	16.7	26.7
50	73.3	0.0	23.3	26.7	60.0	20.0	40.0

L PROSPECTIVE TEST ON PUBLIC MODELS

Twelve public reasoning models (ten planned, two optional ones scored because they finished before the cutoff) and the untouched Qwen3 base model as an anchor (seen under this sampling before registration, so outside every verdict) were decoded on the held-out panel, first planned seed of each item, with each model's recommended sampling (temperature 0.6 and top-p 0.95 except Nemotron-3-Nano, T=1), a 32,768-token budget, the NATIVE contract and, where the template opens the thought in the prompt, PRECLOSED, and scored with the registered judge. Three predictions were frozen before any output of this test existed (preregistration sha256 `35477683...`): P-A, $|\text{artifact}| \le a + b + \hat{\delta}$ with $\hat{\delta}$ fixed in advance at one row or one point; P-B, when $|\text{artifact}| > 0.5\,\text{pp}$, its sign is that of $a - b$; P-C, when $a + b < 2\,\text{pp}$, $|\text{artifact}| \le 2\,\text{pp}$. All three held in every eligible cell (42/42, 14/14 and 16/16; Table 34), as did the judge-free form $|\text{artifact}| \le a + b$ (with the anchor, 45 of 46 cells: its PRECLOSED GPQA-main cell has a=b=0 and one disagreement row, as in the held-out P1 failure). The errors are large: on GPQA-main under NATIVE, R1-Distill-Qwen-7B scores 0.8% strict and 56.0% committed and OpenReasoning-Nemotron-32B 11.2 and 63.6%, because they end with a plain "ANSWER: C" (b close to 1) although the prompt asks for a boxed letter; Nemotron-3-Nano on AIME-2024 gains 13.3 points of truncation credit ($a = 16.7\%$). The ranking comparison was registered as descriptive, over all 13 models under NATIVE: strict and committed rankings nearly agree on AIME (Kendall $\tau_b = 0.90$, 3 inverted pairs, none with both CIs excluding zero) but not on GPQA-main ($\tau_b = 0.64$; 13 inverted pairs, 8 with both the strict and the committed CI excluding zero); over the twelve scored models τ_b is 0.585 (GPQA-main) and 0.880 (AIME-2024), and 6 of the 13 GPQA-main inversions survive simultaneous intervals (Table 20). No inversion involves the anchor, so among the twelve scored models 13 of 66 pairs invert on GPQA-main and 3 of 66 on AIME. This checks the diagnostic and measures a grader's error on deployed models; it is not a ranking of the models, whose answer formats were trained for other harnesses, and each cell holds one decoding per item.

M ADDED STUDENTS

Two distilled reasoning models were trained on both recipes with predictions registered in advance. Both tests use the held-out panel and the registered runner and judge; their evaluation prompts are re-rendered with the respective student's own chat template for each of the two contracts.

M.1 DEEPSEEK-R1-DISTILL-QWEN-7B (R1-QWEN)

R1-Qwen (a dense Qwen2.5-Math-7B model distilled from DeepSeek-R1) was trained with the paper's data files, validation split, seeds and schedule (learning rate 10^{-5}, cosine, batch 128 $\times$ 32,768 packed tokens, three-epoch horizon). Predictions S1–S6 were frozen before any checkpoint of this student existed (sha256 `d6925a56...`); the first training jobs failed at start-up before producing a step and were resubmitted unchanged. Steps 50 and 100 were evaluated as added checkpoints. Table 35 gives the verdicts and Table 36 the cells.

Table 34: Public reasoning models on the held-out panel under NATIVE (predictions P-A to P-C registered before any output; table descriptive; one decoding per item, recommended sampling, registered judge; %). Strict: last-`\boxed{}` rule; comm.: delivered accuracy (judge-read visible answer); a: rows that hit the budget or end unclosed yet carry a strict parse; b: closed rows without one; ^{o}: optional model, scored because it finished before the cutoff; anchor: the untouched Qwen3 base model. The PRECLOSED cells, which also enter the verdicts, are in the released results. GPQA-main: 250 items outside GPQA Diamond; AIME-2024: 30 items.

	GPQA-main				AIME-2024			
model	strict	comm.	a	b	strict	comm.	a	b
Qwen3-30B-A3B-Thinking-2507 (anchor)	67.6	67.6	0.0	0.0	86.7	86.7	0.0	0.0
R1-Distill-Qwen-1.5B	2.8	35.6	0.0	94.4	30.0	30.0	0.0	3.3
R1-Distill-Qwen-7B	0.8	56.0	0.0	98.8	50.0	50.0	0.0	3.3
R1-Distill-Qwen-14B^o	65.6	66.8	0.0	2.0	66.7	66.7	0.0	0.0
R1-Distill-Qwen-32B	66.8	66.8	0.0	0.0	70.0	70.0	0.0	0.0
Light-R1-7B-DS^o	0.0	60.4	0.0	99.6	66.7	66.7	0.0	3.3
Light-R1-14B-DS	66.8	67.6	0.0	1.6	70.0	70.0	0.0	0.0
Light-R1-32B	58.8	57.2	2.8	0.0	76.7	76.7	0.0	0.0
TinyR1-32B-Preview	72.0	72.0	0.0	0.0	80.0	80.0	0.0	0.0
OpenMath-Nemotron-7B	23.6	28.8	0.0	24.8	63.3	63.3	0.0	3.3
OpenReasoning-Nemotron-32B	11.2	63.6	0.0	82.8	86.7	86.7	0.0	0.0
Llama-3.1-Nemotron-Nano-8B	42.4	42.0	0.8	0.0	56.7	56.7	0.0	3.3
Nemotron-3-Nano-30B-A3B	62.8	62.4	2.4	0.0	80.0	66.7	16.7	0.0

R1-Qwen states GPQA letters without a box from the start: under NATIVE its base model scores 0.6% strict and 57.0% delivered on GPQA-main ($b = 99.2\%$), and every GPQA-main gap is negative, as $b > a$ predicts. S1, S5 and S6 hold; S2 has no eligible cell; S3 fails with the opposite sign, because unread NATIVE answers make strict understate NATIVE more than PRECLOSED; S4 fails on GPQA-main. The larger count gives the sign of 29 of the 34 nonzero gaps; for R1-Qwen this is not a registered prediction (the sign row of Table 2). The public-model test decoded the same base model on the same items with its recommended sampling (T=0.6, one decoding per item) and found 0.8% strict and 56.0% committed (Table 34); the figures here use the registered protocol (T=1, all decoding seeds).

Table 35: Registered predictions for DeepSeek-R1-Distill-Qwen-7B (R1-Qwen), frozen before any checkpoint existed. Gap: strict − delivered accuracy (pp); a: unfinished rows with a strict parse; b: finished rows without one; δ: parser–judge disagreement mass.

	prediction	observed	verdict
S1	step-25 budget-hit or unclosed share of NATIVE AIME rows exceeds the base model's by ≥ 20 points	31.7 and 84.2 over base (Gemini, gpt-oss)	PASS
S2	step-25 strict − delivered > 0 where $a > b + 1$	no eligible cell ($a \leq b + 1$ in all eight)	PASS (vacuous)
S3	Gemini step 200 GPQA-main: PRECLOSED−NATIVE strict gap below the delivered gap by ≥ 5	strict $+12.2$, delivered $+1.0$, difference $+11.2$	FAIL
S4	gpt-oss step 200: |strict gap − delivered gap| ≤ 3 on both tasks	AIME-2024 0.0, GPQA-main $+7.6$	FAIL
S5	|strict − delivered| $\leq a + b + \delta$ in every cell	36 of 36 cells	PASS
S6	$a + b < 2 \Rightarrow$ |strict − delivered| ≤ 2	1 of 1 cells	PASS

M.2 DEEPSEEK-R1-DISTILL-LLAMA-8B (R1-LLAMA)

R1-Llama, DeepSeek-R1-Distill-Llama-8B (a dense Llama-3.1-8B model distilled from DeepSeek-R1, the only non-Qwen family in our registered training tests; the Gemma-4-31B runs of Appendix R are exploratory), was trained on both recipes with the paper's data, seeds and schedule (learning rate 10^{-5}, cosine, batch 128, three-epoch horizon) and evaluated on the held-out panel with the registered runner and judge, the prompts re-rendered with its own chat template. Predictions S1–S6 are R1-Qwen's (Appendix M.1), copied verbatim and frozen before either training job was submitted (sha256 `08c3c1c3...`). A1–A3 were added because S5 and S6 hold by construction; they were frozen after the judge had processed the base and step-25 runs but before any of its outputs was opened (only the judge-free counts a and b of those runs, which need no judge, were read to write them; sha256 `ee2c7805...`). Step 100 was evaluated as an added checkpoint and step 50 was skipped; A1 covers every evaluated checkpoint. Table 37 gives the verdicts and Table 38 the cells.

Table 36: DeepSeek-R1-Distill-Qwen-7B (R1-Qwen) on the held-out panel (%, pp; columns as in Table 38). Steps 50 and 100 are additional checkpoints. B: budget hit; U: unclosed thought; gap: strict minus delivered accuracy. Base denotes the untouched student.

recipe	step	task	contract	B	U	strict	delivered	gap	a	b
base		AIME-2024	NATIVE	0.0	0.0	47.5	47.5	0.0	0.0	1.7
base		AIME-2024	PRECLOSED	0.0	0.0	16.7	17.5	-0.8	0.0	8.3
base		GPQA-main	NATIVE	0.0	0.2	0.6	57.0	-56.4	0.0	99.2
base		GPQA-main	PRECLOSED	0.0	0.0	18.2	44.4	-26.2	0.0	57.2
Gemini	25	AIME-2024	NATIVE	4.2	27.5	12.5	11.7	$+0.8$	6.7	20.8
Gemini	25	AIME-2024	PRECLOSED	0.0	0.0	5.8	5.8	0.0	0.0	42.5
Gemini	25	GPQA-main	NATIVE	0.0	5.4	0.4	46.2	-45.8	0.2	93.8
Gemini	25	GPQA-main	PRECLOSED	0.0	0.0	16.4	39.8	-23.4	0.0	68.0
Gemini	50	AIME-2024	NATIVE	41.7	6.7	1.7	6.7	-5.0	2.5	44.2
Gemini	50	AIME-2024	PRECLOSED	35.0	0.0	0.8	10.8	-10.0	3.3	52.5
Gemini	50	GPQA-main	NATIVE	0.0	6.8	1.4	41.4	-40.0	0.0	91.8
Gemini	50	GPQA-main	PRECLOSED	0.2	0.0	2.8	39.2	-36.4	0.0	93.6
Gemini	100	AIME-2024	NATIVE	66.7	1.7	0.8	13.3	-12.5	0.8	28.3
Gemini	100	AIME-2024	PRECLOSED	49.2	0.0	0.8	13.3	-12.5	0.0	47.5
Gemini	100	GPQA-main	NATIVE	1.6	5.8	0.2	40.4	-40.2	0.0	92.4
Gemini	100	GPQA-main	PRECLOSED	0.2	0.0	7.4	43.4	-36.0	0.0	87.6
Gemini	200	AIME-2024	NATIVE	45.8	0.0	0.0	10.8	-10.8	0.0	54.2
Gemini	200	AIME-2024	PRECLOSED	27.5	0.0	2.5	12.5	-10.0	0.0	67.5
Gemini	200	GPQA-main	NATIVE	0.6	3.4	1.0	39.4	-38.4	0.0	95.0
Gemini	200	GPQA-main	PRECLOSED	0.4	0.0	13.2	40.4	-27.2	0.0	75.4
gpt-oss	25	AIME-2024	NATIVE	80.0	4.2	11.7	10.0	$+1.7$	5.0	5.0
gpt-oss	25	AIME-2024	PRECLOSED	2.5	0.0	0.8	3.3	-2.5	0.8	66.7
gpt-oss	25	GPQA-main	NATIVE	17.0	1.4	3.6	34.8	-31.2	0.0	77.2
gpt-oss	25	GPQA-main	PRECLOSED	2.0	0.0	14.4	32.0	-17.6	0.0	60.0
gpt-oss	50	AIME-2024	NATIVE	62.5	0.0	25.8	28.3	-2.5	7.5	6.7
gpt-oss	50	AIME-2024	PRECLOSED	58.3	0.0	17.5	20.0	-2.5	10.8	15.0
gpt-oss	50	GPQA-main	NATIVE	28.2	0.6	7.2	33.6	-26.4	0.4	61.0
gpt-oss	50	GPQA-main	PRECLOSED	17.2	0.0	18.0	35.6	-17.6	1.4	47.6
gpt-oss	100	AIME-2024	NATIVE	62.5	0.0	30.0	30.8	-0.8	5.8	3.3
gpt-oss	100	AIME-2024	PRECLOSED	53.3	0.0	22.5	21.7	$+0.8$	5.8	8.3
gpt-oss	100	GPQA-main	NATIVE	20.6	0.0	14.4	34.4	-20.0	1.2	55.8
gpt-oss	100	GPQA-main	PRECLOSED	12.8	0.0	25.6	35.0	-9.4	0.4	33.2
gpt-oss	200	AIME-2024	NATIVE	60.0	0.0	35.8	31.7	$+4.2$	8.3	3.3
gpt-oss	200	AIME-2024	PRECLOSED	59.2	0.0	25.8	21.7	$+4.2$	11.7	3.3
gpt-oss	200	GPQA-main	NATIVE	25.8	0.2	17.4	33.4	-16.0	1.0	46.2
gpt-oss	200	GPQA-main	PRECLOSED	13.4	0.0	28.4	36.8	-8.4	0.4	24.6

The student reproduces both halves of the Qwen3 pattern: its step-25 Gemini gaps are positive (A2) and every step-200 cell whose count b exceeds a by 5 points has a negative gap (A3). The two contract-level predictions fail. Under the Gemini recipe strict grading misses delivered answers in both contracts (step-200 GPQA-main gaps -13.8 under NATIVE and -15.8 under PRECLOSED), so the PRECLOSED$-$NATIVE difference that S3 tests is small (-2.0), and the gpt-oss recipe shows a contract-level artifact on AIME-2024 (S4, -4.2). S5 was registered with $+\delta$; the judge-free form $|\text{gap}| \le a + b$ also holds in all 28 cells.

Table 37: Registered predictions for DeepSeek-R1-Distill-Llama-8B (S1–S6: frozen before training; A1–A3: addendum frozen before any judgement was read). Gap: strict $-$ delivered accuracy (pp); a: unfinished rows with a strict parse; b: finished rows without one; δ: parser–judge disagreement mass.

	prediction	observed	verdict
S1	step-25 budget-hit or unclosed share of NATIVE AIME rows exceeds the base model's by ≥ 20 points	90.0 and 70.8 over base (Gemini, gpt-oss)	PASS
S2	step-25 strict $-$ delivered > 0 where $a > b + 1$	4 of 4 cells, $+0.8$ to $+11.6$	PASS
S3	Gemini step 200 GPQA-main: PRECLOSED$-$NATIVE strict gap below the delivered gap by ≥ 5	strict -1.4, delivered $+0.6$, difference -2.0	FAIL
S4	gpt-oss step 200: $\lvert$strict gap $-$ delivered gap$\rvert \le 3$ on both tasks	AIME-2024 -4.2, GPQA-main $+1.4$	FAIL
S5	$\lvert$strict $-$ delivered$\rvert \le a + b + \delta$ in every cell	28 of 28 cells	PASS
S6	$a + b < 2 \Rightarrow \lvert$strict $-$ delivered$\rvert \le 2$	1 of 1 cells	PASS
A1	sign of the gap = sign of $a - b$ where $\max(a, b) \ge 5$ and $\lvert a - b \rvert \ge 5$ ($\ge 90\%$)	20 of 20 cells	PASS
A2	Gemini step-25 NATIVE gap > 0 on both tasks	AIME-2024 $+2.5$, GPQA-main $+11.6$	PASS
A3	step-200 gap < 0 wherever $b - a \ge 5$	4 of 4 cells, -4.2 to -15.8	PASS

Table 38: DeepSeek-R1-Distill-Llama-8B on the held-out panel (%, pp; AIME-2024: 30 items × 4 seeds; GPQA-main: 250 items × 2 seeds; T=1). B: budget hit; U: ended inside an unclosed thought; delivered: the registered judge's reading of the visible answer, nothing for B and U rows; gap: strict − delivered; a: B or U rows with a strict parse; b: closed rows without one. NATIVE/PRECLOSED: generation starts inside / after the thought. Step 100: added.

recipe	step	task	contract	B	U	strict	delivered	gap	a	b
base		AIME-2024	NATIVE	0.0	0.0	37.5	37.5	0.0	0.0	4.2
base		AIME-2024	PRECLOSED	0.0	0.0	41.7	41.7	0.0	0.0	3.3
base		GPQA-main	NATIVE	0.0	0.2	49.4	55.0	−5.6	0.0	9.4
base		GPQA-main	PRECLOSED	0.0	0.0	49.2	53.4	−4.2	0.0	9.2
Gemini	25	AIME-2024	NATIVE	21.7	68.3	5.8	3.3	+2.5	38.3	2.5
Gemini	25	AIME-2024	PRECLOSED	40.8	0.0	4.2	5.8	−1.7	16.7	34.2
Gemini	25	GPQA-main	NATIVE	11.8	55.0	24.2	12.6	+11.6	48.0	15.2
Gemini	25	GPQA-main	PRECLOSED	12.8	0.0	27.6	37.0	−9.4	11.0	30.2
Gemini	100	AIME-2024	NATIVE	42.5	0.0	0.0	1.7	−1.7	0.0	56.7
Gemini	100	AIME-2024	PRECLOSED	47.5	0.0	0.0	1.7	−1.7	0.0	51.7
Gemini	100	GPQA-main	NATIVE	2.2	2.6	16.6	30.4	−13.8	1.0	49.0
Gemini	100	GPQA-main	PRECLOSED	1.0	0.0	17.0	32.8	−15.8	0.0	49.4
Gemini	200	AIME-2024	NATIVE	31.7	0.0	0.0	4.2	−4.2	0.0	66.7
Gemini	200	AIME-2024	PRECLOSED	23.3	0.0	0.0	5.0	−5.0	0.0	76.7
Gemini	200	GPQA-main	NATIVE	0.6	1.8	18.6	32.4	−13.8	0.2	40.0
Gemini	200	GPQA-main	PRECLOSED	0.8	0.0	17.2	33.0	−15.8	0.0	47.0
gpt-oss	25	AIME-2024	NATIVE	68.3	2.5	18.3	17.5	+0.8	12.5	3.3
gpt-oss	25	AIME-2024	PRECLOSED	69.2	0.0	17.5	15.8	+1.7	11.7	5.8
gpt-oss	25	GPQA-main	NATIVE	22.6	0.6	27.0	36.4	−9.4	4.0	22.6
gpt-oss	25	GPQA-main	PRECLOSED	24.2	0.0	26.4	34.2	−7.8	6.0	22.8
gpt-oss	100	AIME-2024	NATIVE	53.3	0.0	30.0	29.2	+0.8	4.2	6.7
gpt-oss	100	AIME-2024	PRECLOSED	63.3	0.0	28.3	28.3	0.0	0.8	0.8
gpt-oss	100	GPQA-main	NATIVE	24.8	0.2	33.0	35.6	−2.6	6.6	13.4
gpt-oss	100	GPQA-main	PRECLOSED	25.2	0.0	31.4	35.8	−4.4	6.6	14.6
gpt-oss	200	AIME-2024	NATIVE	50.0	0.0	38.3	35.8	+2.5	6.7	5.0
gpt-oss	200	AIME-2024	PRECLOSED	48.3	0.0	35.8	37.5	−1.7	1.7	2.5
gpt-oss	200	GPQA-main	NATIVE	22.2	0.2	38.2	39.2	−1.0	6.0	7.6
gpt-oss	200	GPQA-main	PRECLOSED	21.4	0.0	41.6	41.2	+0.4	7.6	7.6

N A CONTROLLED FORMAT TEST ON QWEN3-4B-BASE

Design. Qwen3-4B-Base is a pretrained model without reasoning post-training, so SFT should raise its accuracy. It was trained on three arms with the paper's recipe (learning rate 10^{-5}, cosine to 10^{-6}, batch 128 × 32,768 packed tokens, three-epoch horizon, seed 42, the main Qwen3 student's template): the Gemini recipe, the gpt-oss recipe, and the gpt-oss recipe with every `\boxed{X}` replaced by X (531,909 boxes; 3,431 bare `\boxed` tokens and 779 unbalanced openers removed, their text kept; prompts and row order unchanged). Over-long rows were dropped whole, never truncated (Gemini 5,334, gpt-oss 16,162, unboxed 16,149 of 100,000). The base model and steps 25 and 200 of each arm were evaluated on the held-out panel with the registered judge. Predictions Q1–Q4 were frozen before any training job was submitted (sha256 `fe90c6eb...`). Table 39 gives the verdicts and Table 40 the cells.

Deviations. (i) Layout: the 8-GPU jobs did not schedule, so each arm ran on one node with 4 GPUs (the unboxed arm also on a 2-GPU copy) and the same global batch; before any unboxed checkpoint existed we fixed that the 4-GPU copy's checkpoints would be evaluated, and they were. (ii) Engine length: the model's window is 32,768 positions. The first evaluation let sequences run past it and crashed when a response reached the window, so every run was restarted with a per-request budget of 32,768 minus the prompt (at most 1,055 tokens, median 208); the partial outputs were set aside and not analysed.

Outcome. Q1 passes: SFT raises the gpt-oss student's delivered accuracy (under NATIVE the base model never closes its thought and delivers nothing), which makes Q1 trivially satisfied; the informative, unregistered PRECLOSED comparison is 19.2 vs 5.0% (AIME-2024) and 34.4 vs 32.2% (GPQA-main). Q2 fails through its second clause only: the unboxed student's step-200 PRECLOSED gaps are −18.3 and −21.6 points, while the boxed student's GPQA-main gap is −3.8, outside the ±3 registered for it. Q3 passes (25 of 27 eligible cells). Q4 does not apply: the base model already never finishes under NATIVE, so the step-25 rise it conditions on is zero. The Qwen3-4B-Base cells are not in the 238-cell pool of Figure 3. A second training seed (271828) of the two gpt-oss arms was

trained with the same recipe, schedule, data and global batch, with predictions Q5–Q7 (called R1–R3 in the frozen addendum) frozen before any of its jobs was submitted (sha256 `e6a168cc...`); only its step-200 checkpoints were evaluated, with the same runner, panel and judge. All three pass (Table 39): the unboxed student's PRECLOSED gaps are -21.7 and -22.8 points (Q5), 23.3 and 20.8 points below the boxed student's (Q6), and strict accuracy ranks it below the boxed student on both tasks while its delivered accuracy is higher on both (Q7). The boxed student's GPQA-main gap is -2.0, inside the ± 3 band that Q2 set for seed 42. Averaged over the two seeds, the PRECLOSED gaps are -20.0 and -22.2 points (unboxed) and $+1.7$ and -2.9 (boxed) on AIME-2024 and GPQA-main; Table 41 gives the seed-271828 cells.

Table 39: Registered predictions for Qwen3-4B-Base, frozen before training (Q5–Q7: second training seed). Unboxed: the gpt-oss targets with every box replaced by its content. Gap: strict − delivered accuracy (pp); a, b as in Table 37.

	prediction	observed	verdict
Q1	gpt-oss step 200 NATIVE delivered accuracy above the base model's on both tasks	AIME-2024 15.0 vs 0.0, GPQA-main 20.6 vs 0.0	PASS
Q2	step 200 PRECLOSED: unboxed gap ≤ -5 on both tasks and gpt-oss $\lvert\text{gap}\rvert \leq 3$ on both	unboxed -18.3 / -21.6; gpt-oss $+1.7$ / -3.8	FAIL
Q3	sign of the gap = sign of $a - b$ where $\max(a, b) \geq 5$ and $\lvert a - b\rvert \geq 5$ ($\geq 90\%$)	25 of 27 cells	PASS
Q4	if the step-25 NATIVE AIME budget-hit or unclosed share rises by ≥ 20 points over base: gap > 0	rise 0.0 (Gemini), 0.0 (gpt-oss)	not applicable
Q5	seed 271828, step 200 PRECLOSED: unboxed gap ≤ -5 on both tasks	unboxed -21.7 / -22.8	PASS
Q6	seed 271828: unboxed gap at least 10 points below the gpt-oss gap on both tasks	-23.3 / -20.8	PASS
Q7	seed 271828: unboxed strict below gpt-oss strict on both tasks; unboxed delivered not more than 5 below	strict 0.0 vs 20.0 / 8.8 vs 26.0; delivered 21.7 vs 18.3 / 31.6 vs 28.0	PASS

Table 40: Qwen3-4B-Base on the held-out panel (%, pp; columns as in Table 38); gap with its item-cluster 95% CI. Unboxed: the gpt-oss targets with every box replaced by its content.

recipe	step	task	contract	B	U	strict	delivered	gap	a	b
base		AIME-2024	NATIVE	0.0	100.0	5.8	0.0	$+5.8$ [$+1.7$, $+11.7$]	55.0	0.0
base		AIME-2024	PRECLOSED	0.8	0.0	5.0	5.0	0.0 [0.0, 0.0]	0.0	16.7
base		GPQA-main	NATIVE	1.2	98.8	11.4	0.0	$+11.4$ [$+8.4$, $+14.4$]	39.4	0.0
base		GPQA-main	PRECLOSED	0.2	0.0	31.6	32.2	-0.6 [-2.2, $+1.2$]	0.2	11.4
Gemini	25	AIME-2024	NATIVE	79.2	20.8	1.7	0.0	$+1.7$ [0.0, $+4.2$]	32.5	0.0
Gemini	25	AIME-2024	PRECLOSED	78.3	0.0	0.8	0.8	0.0 [-2.5, $+2.5$]	22.5	12.5
Gemini	25	GPQA-main	NATIVE	71.4	28.6	24.0	0.0	$+24.0$ [$+20.2$, $+28.0$]	77.8	0.0
Gemini	25	GPQA-main	PRECLOSED	64.4	0.0	18.6	7.8	$+10.8$ [$+7.6$, $+14.0$]	42.4	6.8
Gemini	200	AIME-2024	NATIVE	70.8	6.7	0.0	5.8	-5.8 [-12.5, -0.8]	1.7	22.5
Gemini	200	AIME-2024	PRECLOSED	63.3	0.0	0.0	6.7	-6.7 [-13.3, -1.7]	0.8	34.2
Gemini	200	GPQA-main	NATIVE	4.4	30.8	28.8	22.8	$+6.0$ [$+3.4$, $+8.6$]	26.8	5.6
Gemini	200	GPQA-main	PRECLOSED	3.2	0.0	22.8	27.0	-4.2 [-6.2, -2.4]	0.2	13.2
gpt-oss	25	AIME-2024	NATIVE	99.2	0.8	7.5	0.0	$+7.5$ [$+0.8$, $+15.8$]	49.2	0.0
gpt-oss	25	AIME-2024	PRECLOSED	95.8	0.0	9.2	2.5	$+6.7$ [$+0.8$, $+14.2$]	35.8	0.8
gpt-oss	25	GPQA-main	NATIVE	93.2	6.8	17.6	0.0	$+17.6$ [$+14.2$, $+21.2$]	61.8	0.0
gpt-oss	25	GPQA-main	PRECLOSED	91.4	0.0	20.2	0.4	$+19.8$ [$+16.2$, $+23.4$]	56.4	1.8
gpt-oss	200	AIME-2024	NATIVE	74.2	4.2	19.2	15.0	$+4.2$ [$+0.8$, $+7.5$]	14.2	2.5
gpt-oss	200	AIME-2024	PRECLOSED	65.8	0.0	20.8	19.2	$+1.7$ [-1.7, $+5.0$]	10.8	4.2
gpt-oss	200	GPQA-main	NATIVE	24.4	21.8	26.8	20.6	$+6.2$ [$+3.0$, $+9.4$]	25.8	12.2
gpt-oss	200	GPQA-main	PRECLOSED	23.8	0.0	30.6	34.4	-3.8 [-5.6, -2.0]	1.8	10.2
gpt-oss, unboxed	25	AIME-2024	NATIVE	94.2	5.8	2.5	0.0	$+2.5$ [0.0, $+5.8$]	5.8	0.0
gpt-oss, unboxed	25	AIME-2024	PRECLOSED	93.3	0.0	2.5	0.8	$+1.7$ [0.0, $+4.2$]	3.3	5.8
gpt-oss, unboxed	25	GPQA-main	NATIVE	88.2	11.8	10.0	0.0	$+10.0$ [$+7.2$, $+13.0$]	36.8	0.0
gpt-oss, unboxed	25	GPQA-main	PRECLOSED	90.4	0.0	12.0	1.4	$+10.6$ [$+7.4$, $+14.0$]	32.2	5.2
gpt-oss, unboxed	200	AIME-2024	NATIVE	70.8	3.3	0.0	20.8	-20.8 [-33.3, -10.0]	0.0	25.8
gpt-oss, unboxed	200	AIME-2024	PRECLOSED	73.3	0.0	0.0	18.3	-18.3 [-30.0, -8.3]	0.0	26.7
gpt-oss, unboxed	200	GPQA-main	NATIVE	32.0	15.4	6.0	23.4	-17.4 [-22.0, -13.0]	9.4	47.4
gpt-oss, unboxed	200	GPQA-main	PRECLOSED	33.8	0.0	6.0	27.6	-21.6 [-26.0, -17.2]	1.0	53.8

O SEEDS, LEARNING RATES AND DATA IN THE SECOND HARNESS

Three training seeds per recipe (42, 271828, 314159) were evaluated on the fixed grid of steps 25–200 (eight checkpoints) under a 65k-token budget with four decoding runs, native contract

Table 41: Qwen3-4B-Base, second training seed (271828), step 200 on the held-out panel (%, pp; columns as in Table 40); gap with its item-cluster 95% CI.

recipe	step	task	contract	B	U	strict	delivered	gap	a	b
gpt-oss	200	AIME-2024	NATIVE	75.8	6.7	19.2	10.8	$+8.3$ $[+3.3, +15.0]$	15.0	2.5
gpt-oss	200	AIME-2024	PRECLOSED	70.0	0.0	20.0	18.3	$+1.7$ $[0.0, +4.2]$	7.5	1.7
gpt-oss	200	GPQA-main	NATIVE	22.8	16.0	26.0	21.6	$+4.4$ $[+1.8, +7.0]$	19.2	12.0
gpt-oss	200	GPQA-main	PRECLOSED	20.6	0.0	26.0	28.0	-2.0 $[-3.6, -0.6]$	2.6	12.2
gpt-oss, unboxed	200	AIME-2024	NATIVE	68.3	6.7	0.0	15.0	-15.0 $[-25.0, -6.7]$	0.0	25.0
gpt-oss, unboxed	200	AIME-2024	PRECLOSED	67.5	0.0	0.0	21.7	-21.7 $[-32.5, -11.7]$	0.0	32.5
gpt-oss, unboxed	200	GPQA-main	NATIVE	24.2	22.4	6.8	23.0	-16.2 $[-20.4, -12.0]$	13.2	47.8
gpt-oss, unboxed	200	GPQA-main	PRECLOSED	21.6	0.0	8.8	31.6	-22.8 $[-27.0, -18.6]$	2.2	61.8

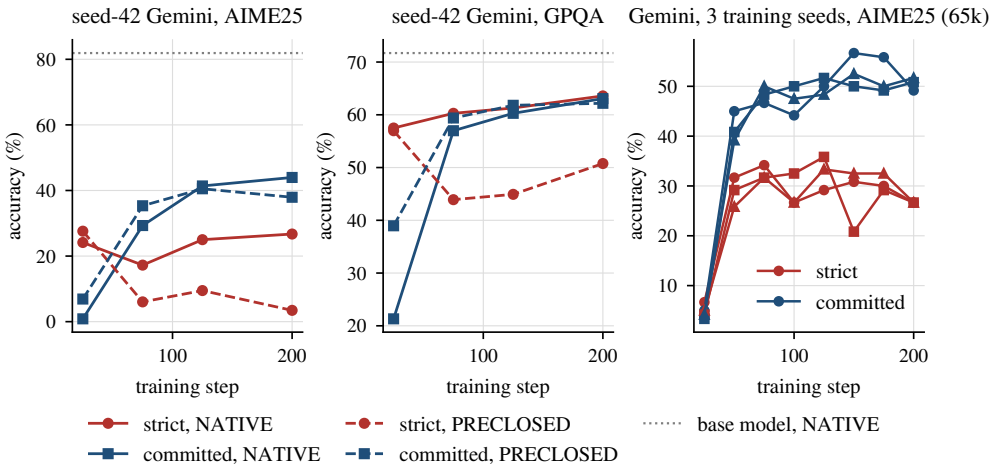

Figure 11: Learning curves under strict (red) and delivered-answer (blue) scoring. Left, centre: seed-42 Gemini under both contracts (32k harness; solid NATIVE, dashed PRECLOSED), with the untouched base model (NATIVE) as a dotted anchor. Right: Gemini on AIME across three training seeds under the 65k harness. At 32k the strict curves start from a false floor far above delivered accuracy and dip in three of the four arms while delivered accuracy rises; at 65k both start near zero, and the committed curves stay close across seeds where the strict curves dip and diverge. Strict: the frozen rule reading the last `\boxed{}` answer after the last thought closer (else the whole output); committed: the final answer the visible answer channel states, read by a gold-blind judge; NATIVE/PRECLOSED: the prompt contracts in which generation starts inside the thought / after an appended empty thought closer; 32k, 65k: token budgets of the discovery harness and the second harness; arm: one contract; Gemini: the recipe distilled from a Gemini-family teacher, named for its teacher.

only, and scored by nemo-skills' extractor on the post-thought text (43,776 rows; 6.0% hit the budget). The harness records the post-thought text separately, so a response that ended inside an unclosed thought leaves that field empty; 1,818 terminated rows have an empty answer field, 1,786 of them Gemini and 931 of the judge-readable ones at step 25, and they are scored uncommitted, consistent with §4 (reading them on the whole output instead gives the whole-output figures quoted in parentheses). Table 42 gives the Gemini AIME curves. Under this harness truncation credit on AIME is confined to step 25, where budget-hit rows that had closed the thought add 0.8–1.7 points to Gemini and 8.3–14.2 to gpt-oss, but the format artifact remains: across three training seeds the Gemini AIME strict curve sits at 21–36% from step 50 while the committed curve sits at 39–57%, because the model commits a visible answer on 93–99% of rows from step 100 but the extractor parses only 46–68%. One seed's 15-point strict dip at step 150 (35.8 $\rightarrow$ 20.8%) is a parse-rate fluctuation (committed 51.7 $\rightarrow$ 50.0%). The committed curves are not monotone either (after step 75 the 30-item, four-seed panel moves by up to 6.7 points between adjacent checkpoints), but they rise from a near-zero floor and stay within a few points of each other across seeds, where the strict curves dip and diverge (Figure 11). The recipe gap that motivated the program—Gemini below gpt-oss by 28.75 pp of balanced early-grid AUC—is 19.55 pp under delivered scoring (16.70

if unclosed thoughts are read on their whole output); in signed gpt-oss−Gemini terms, on GPQA it is real (+8.6 pp strict and committed at step 200), on AIME half of it was format (+51.4 pp strict, +27.5 pp committed). Early-grid AUC is the trapezoidal area under the accuracy curve over the eight checkpoints of steps 25–200 divided by the step range (175 steps), i.e. the mean of the eight accuracies with the first and last at half weight; "balanced" averages AIME and GPQA with equal weight, so GPQA's 198 items do not outweigh AIME's 30. This strict summary is the family's pre-registered primary accuracy estimand; its pre-registered exact test over the 20 balanced allocations of the six trajectories reaches its floor, $p = 0.1$ (Holm-adjusted 0.3 over the three primary metrics), because every Gemini trajectory lies below every gpt-oss trajectory. The delivered-answer re-scoring of this family is post hoc. Pairing each Gemini trajectory with the gpt-oss trajectory of the same training seed, on balanced early-grid AUC gpt-oss minus Gemini is +28.75 pp strict (per seed +29.75, +27.91, +28.58) and +19.55 pp committed (+20.13, +18.77, +19.75; whole-output reading +16.70); at step 200, GPQA +8.63 strict and +8.59 pp committed, AIME +51.39 and +27.50 pp. The gpt-oss student's GPQA strict accuracy at step 25 (58–60%) exceeds its delivered accuracy (49–51%): this extractor also awards credit to some budget-hit rows.

Table 42: Second harness: Gemini on AIME25 by training seed (%). Rows: strict accuracy / visible-channel delivered accuracy / commitment rate / strict parse rate. Second harness: 65k-token budget, its own strict extractor (nemo-skills), NATIVE contract only; NATIVE: generation starts inside the thought; committed: the final answer the visible answer channel states, read by a gold-blind judge; commit rate: share of rows with a delivered answer; parse rate: share of rows with a strict parse; step: optimizer step; Gemini: the recipe distilled from a Gemini-family teacher, named for its teacher. AIME25: AIME 2025.

		step							
seed	metric	25	50	75	100	125	150	175	200
42	strict	6.7	31.7	34.2	26.7	29.2	30.8	30.0	26.7
	committed	5.0	45.0	46.7	44.2	50.0	56.7	55.8	49.2
	commit rate	5.0	70.8	82.5	95.0	95.8	99.2	96.7	96.7
	parse rate	6.7	50.8	60.8	58.3	61.7	60.8	50.0	52.5
271828	strict	4.2	29.2	31.7	32.5	35.8	20.8	29.2	26.7
	committed	3.3	40.8	48.3	50.0	51.7	50.0	49.2	50.8
	commit rate	4.2	68.3	86.7	95.8	93.3	98.3	98.3	98.3
	parse rate	5.0	47.5	55.8	66.7	65.8	46.7	60.8	45.8
314159	strict	5.0	25.8	31.7	26.7	33.3	32.5	32.5	26.7
	committed	4.2	39.2	50.0	47.5	48.3	52.5	50.0	51.7
	commit rate	5.0	76.7	87.5	95.8	96.7	99.2	98.3	98.3
	parse rate	5.8	51.7	56.7	61.7	67.5	63.3	62.5	53.3

O.1 Learning rate (exploratory, post hoc)

The Gemini recipe was also trained at peak learning rates $5 \cdot 10^{-6}$ and $2 \cdot 10^{-6}$ (decaying to a tenth of the peak; otherwise as in Appendix B) and scored on AIME 2025 in this harness with the same judge, at checkpoints that differ by rate (Table 43; the 10^{-5} run is seed 42 of Table 42, whose step-200 checkpoint was evaluated separately). Strict and delivered accuracy both bottom out at step 25, 50 and 100 respectively (at 10^{-5} the first checkpoint evaluated, so no descent is observed) and differ little there (6.7 vs. 5.0, 8.3 vs. 10.0, 13.3 vs. 16.7%): budget-hit and unclosed rows earn almost no credit in this harness (strict credit in two of the 27 cells, 1.7 points each, both at step 25). The valley's depth is the non-termination share: at the bottoms 95.0, 80.8 and 66.7% of rows hit the budget or end inside an unclosed thought, while the completed rows are committed-correct on 6 of 6, 52.2 and 50.0%, against 44.9–63.5% at every later checkpoint. The format gap runs on a separate clock: delivered accuracy first exceeds strict by more than 10 points at step 50, 150 and 375, and from there by 12.5–25.8, 12.5–23.3 and 9.2–15.8 points, not monotonically, with item-cluster 95% intervals that exclude zero. A second strict dip at 10^{-5}, step 100 (34.2 → 26.7%, with 5.0% of rows unterminated), is mostly format: delivered accuracy moves 46.7 → 44.2%. Both clocks run at every rate, later at lower ones.

O.2 Learning rate on the held-out panel (exploratory, post hoc)

Five lower-rate checkpoints of Table 43, at and after each rate's valley bottom (at $2 \cdot 10^{-6}$ also before it), were also decoded on the held-out panel with the registered protocol (Appendix J; T=1, both

Table 43: Learning-rate sweep: Gemini on AIME 2025 under the 65k harness, training seed 42 at each peak learning rate (%; 120 rows per cell, 30 items × 4 runs). Strict: the harness's verdict; comm.: committed (visible channel; budget-hit and unclosed rows commit to nothing); B+U: rows that hit the budget or end inside an unclosed thought. Bold: each rate's strict minimum; †: first checkpoint at which comm. exceeds strict by more than 10 points; —: not in this sweep. Exploratory, post hoc. 65k harness: the second harness (65k-token budget, nemo-skills strict extractor); delivered answers are read by a gold-blind judge; peak: peak learning rate; step: optimizer step; Gemini: the recipe distilled from a Gemini-family teacher, named for its teacher.

	peak 10^{-5}			peak $5 \cdot 10^{-6}$			peak $2 \cdot 10^{-6}$		
step	strict	comm.	B+U	strict	comm.	B+U	strict	comm.	B+U
25	**6.7**	5.0	95.0	50.0	48.3	45.0	85.0	85.0	0.0
50	31.7	45.0†	29.2	**8.3**	10.0	80.8	56.7	59.2	28.3
75	34.2	46.7	17.5	30.8	39.2	30.0	28.3	29.2	55.0
100	26.7	44.2	5.0	35.8	42.5	18.3	**13.3**	16.7	66.7
125	29.2	50.0	4.2	—	—	—	29.2	32.5	42.5
150	30.8	56.7	0.0	36.7	49.2†	5.0	—	—	—
175	30.0	55.8	2.5	—	—	—	33.3	34.2	25.0
200	—	—	—	30.8	47.5	0.8	—	—	—
250	—	—	—	—	—	—	32.5	40.0	10.8
375	—	—	—	—	—	—	36.7	47.5†	4.2
500	—	—	—	—	—	—	34.2	47.5	1.7
625	—	—	—	—	—	—	36.7	45.8	0.8
750	—	—	—	—	—	—	31.7	45.8	2.5
875	—	—	—	—	—	—	40.0	51.7	0.8
885	—	—	—	32.5	55.8	0.8	33.3	49.2	0.0

contracts) and scored with the registered judge (Table 44). Both clocks recur on the new items, later at lower rates. At the valley bottoms (step 50 at $5 \cdot 10^{-6}$, step 100 at $2 \cdot 10^{-6}$) 90.8 and 79.2% of NATIVE AIME-2024 rows hit the budget or end unclosed, the NATIVE arm ends inside its thought on 51.6 and 17.6% of GPQA-main rows, and PRECLOSED is ahead under delivered scoring ($+37.2$ and $+10.2$ pp) while the strict grader puts it behind (-6.4 and -6.2 pp). Later (steps 200 and 375) the PRECLOSED arm's prose answers go unread (artifacts -17.4 and -10.6 pp on GPQA-main). Step 25 at $2 \cdot 10^{-6}$ precedes the valley and behaves like the base model (budget hits at most 9.2%, strict equal to committed in every cell).

Table 44: Learning-rate checkpoints on the held-out panel (exploratory, post hoc; T=1, both contracts, registered judge; %, pp). B+U: share of NATIVE rows that hit the budget or end inside an unclosed thought (AIME: AIME-2024, 30 items; GPQA: GPQA-main, 250 items). Δ: PRECLOSED−NATIVE on GPQA-main under strict and delivered scoring; artifact: strict minus committed Δ, with its item-cluster 95% CI in grey (unadjusted). NATIVE/PRECLOSED: generation starts inside / after the thought; peak: peak learning rate of the Gemini recipe; pp: percentage points.

		B+U, NATIVE		GPQA-main			AIME-2024
peak	step	AIME	GPQA	strict Δ	comm. Δ	artifact	artifact
$5 \cdot 10^{-6}$	50	90.8	72.2	-6.4	$+37.2$	-43.6 [-49.6, -37.8]	-23.3 [-35.0, -12.5]
	200	45.8	5.4	-15.4	$+2.0$	-17.4 [-21.2, -13.8]	-17.5 [-30.0, -5.8]
$2 \cdot 10^{-6}$	25	9.2	0.0	0.0	0.0	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
	100	79.2	29.8	-6.2	$+10.2$	-16.4 [-20.2, -12.8]	-21.7 [-34.2, -10.0]
	375	39.2	7.0	-11.4	-0.8	-10.6 [-13.4, -7.8]	-21.7 [-34.2, -10.0]

O.3 LAST CHECKPOINTS AND THE TRAINING-TIME SYSTEM PROMPT ON THE HELD-OUT PANEL (EXPLORATORY, POST HOC)

At the studied learning rate 10^{-5}, the last checkpoints of the Gemini and gpt-oss students (steps 885 and 702) and their step-25 checkpoints decoded with the training-time system prompt were evaluated on the held-out panel with the registered protocol and judge (Table 45; the registered step-25 and step-200 cells are repeated for reference). At step 885 strict still understates the Gemini student on AIME-2024 by 21.7 (NATIVE) and 29.2 points (PRECLOSED), while its GPQA-main gaps are small ($+0.4$ and -3.2); the gpt-oss student's last-checkpoint gaps are at most 5.2 points in size. With

the system prompt the Gemini student's step-25 NATIVE credit remains (+28.3 on AIME-2024 and +39.4 on GPQA-main).

Table 45: Last checkpoints and the training-time system prompt on the held-out panel (exploratory, post hoc; peak learning rate 10^{-5}, training seed 42; AIME-2024: 30 items × 4 seeds, GPQA-main: 250 items × 2 seeds; T=1; %, pp). Registered: the registered decoding without a system turn; + system prompt: the training-time system string added at evaluation. Gap: strict − committed.

				NATIVE			PRECLOSED		
recipe	step	decoding	task	strict	committed	gap	strict	committed	gap
Gemini	25	registered	AIME-2024	22.5	3.3	+19.2	29.2	9.2	+20.0
Gemini	25	registered	GPQA-main	53.4	23.0	+30.4	52.8	40.8	+12.0
Gemini	25	+ system prompt	AIME-2024	30.0	1.7	+28.3	23.3	14.2	+9.2
Gemini	25	+ system prompt	GPQA-main	51.0	11.6	+39.4	40.4	42.8	−2.4
Gemini	200	registered	AIME-2024	40.8	58.3	−17.5	10.8	52.5	−41.7
Gemini	200	registered	GPQA-main	56.6	56.2	+0.4	45.0	56.2	−11.2
Gemini	885 (last)	registered	AIME-2024	41.7	63.3	−21.7	35.8	65.0	−29.2
Gemini	885 (last)	registered	GPQA-main	58.2	57.8	+0.4	55.0	58.2	−3.2
gpt-oss	25	registered	AIME-2024	36.7	24.2	+12.5	39.2	24.2	+15.0
gpt-oss	25	registered	GPQA-main	53.6	40.6	+13.0	51.4	34.6	+16.8
gpt-oss	25	+ system prompt	AIME-2024	48.3	44.2	+4.2	46.7	44.2	+2.5
gpt-oss	25	+ system prompt	GPQA-main	53.6	47.8	+5.8	53.6	50.6	+3.0
gpt-oss	200	registered	AIME-2024	79.2	79.2	0.0	80.0	80.0	0.0
gpt-oss	200	registered	GPQA-main	56.6	61.8	−5.2	59.8	62.2	−2.4
gpt-oss	702 (last)	registered	AIME-2024	80.0	79.2	+0.8	80.8	82.5	−1.7
gpt-oss	702 (last)	registered	GPQA-main	58.0	61.4	−3.4	57.6	62.8	−5.2

O.4 TRAINING DATA (EXPLORATORY, POST HOC)

Two further trajectories of the same student, run earlier in this harness (AIME 2025, four runs, NATIVE) and judged now with the same judge, change only the training data (peak learning rate 10^{-5}, one training seed each). *Self-distillation* trains on 25,740 of the student's own responses, sampled at T=1 with a token cap and kept without a correctness or closure filter, to questions drawn from the Gemini training file; *mixed* trains on the Gemini and gpt-oss targets together (200,000 rows, each question once per teacher). At step 25 the self-distilled student hits the budget on 2.5% of rows and scores 81.7% under both scorers (base: 83.3%), although by then it has trained on 22% of an epoch (about 5.8k rows, against 7.9k and 8.9k) at a larger cumulative learning rate than the Gemini and gpt-oss students, whose rows are budget-hit or unclosed on 95.0 and 61.7% (mixed: 76.7%). Its strict and committed accuracies coincide at every checkpoint up to step 333 (three epochs), and it keeps terminating although 22% of its targets never close the thought. The mixed student's delivered accuracy exceeds strict by 12.5–17.5 points from step 75 on (every interval excludes zero), against 12.5–25.8 for Gemini from step 50 and at most 0.8 for any gpt-oss seed. The early collapse thus accompanies another model's targets rather than long or unclosed ones, and the late artifact follows the Gemini answer form, diluted but not removed by mixing. The self-distilled targets also differ in form (they begin inside the thought), each run is one seed, and data amount is confounded with data type here.

P THE UNTOUCHED BASE MODEL AND A SYSTEM-PROMPT VARIANT

The base model (Qwen3-30B-A3B-Thinking-2507, no SFT) was run on the frozen seed-42 prompts, shards and sampling parameters (tensor-parallel 4, CUDA graphs enabled; recorded as a labelled anchor, not a pre-registered family). Strict accuracy / strict parse rate / budget-hit rate / share of rows containing the thought closer, on the confirmatory rows: AIME NATIVE 81.9 / 86.2 / 12.9 / 90.5%; AIME PRECLOSED 77.6 / 84.5 / 15.5 / 83.6%; GPQA NATIVE 71.7 / 100.0 / 0.0 / 100.0%; GPQA PRECLOSED 72.1 / 100.0 / 0.0 / 100.0%. Delivered accuracy equals strict accuracy in all four cells, and the PRECLOSED−NATIVE contrasts coincide at −4.3 pp (AIME) and +0.4 pp (GPQA): every terminated base-model response boxes its answer (strict parse rates AIME 86 and 84%, GPQA 100 and 100%; the one terminated AIME row without a parse boxes the non-integer $487/3$), no NATIVE response ends inside an unclosed thought, and the 15–18 AIME budget hits per arm are long, non-repetitive derivations on the hardest items (minimum 8-gram distinctness 0.80). Under the PRECLOSED contract the base model begins "Okay, let's try to solve this problem step by step", reasons at length (median 16k tokens on AIME), re-emits the closer itself (84–100% of rows) and

then boxes, so the two contracts are near-equivalent for it. Both distillation recipes lower these levels substantially within 200 steps (the "net" rows of Table 11 and the step 200 rows of Table 12; Gemini curves in Figure 11). Our AIME figure (81.9% at a 32k budget, temperature 1.0) is consistent with the model card's 85.0% at an 81,920 budget. Adding the training-time system prompt ("You are a helpful assistant.") to the evaluation prompt of the seed-42 Gemini checkpoint at step 200 leaves the format habit intact: GPQA PRECLOSED strict 46.4 vs committed 60.8% (22.7% committed without a strict parse), AIME PRECLOSED 4.3 vs 40.5%; NATIVE GPQA 63.5 vs 63.8%.

Engine-configuration noise. We re-ran the base anchor with the frozen seed-42 runner's exact engine settings (tensor-parallel 8, CUDA graphs disabled, 16 concurrent sequences, 82% memory), same prompts, shards, seeds and sampling. Sampling at temperature 1.0 is not bitwise reproducible across kernels: the same row ends with the same strict outcome on only 85–90% of rows. Cell-level strict accuracy moves by -0.1 and $-1.3\,\mathrm{pp}$ on GPQA (item-cluster CIs $[-2.2, +2.0]$ and $[-4.1, +1.4]$) and by -5.2 and $+4.3\,\mathrm{pp}$ on AIME ($[-10.3, -0.9]$ and $[-1.7, +11.2]$), and the AIME contract contrast changes sign, $-4.3 \rightarrow +5.2\,\mathrm{pp}$. In this one comparison a 29-item AIME cell moved by up to $5.2\,\mathrm{pp}$ with the engine configuration alone, the scale of the residual Gemini contract effect in §5; GPQA cells moved by up to $1.3\,\mathrm{pp}$. The seed-42 Gemini strict AIME dip from step 25 to step 75 (28 to 20 of 116 rows, $-6.9\,\mathrm{pp}$, item-cluster CI $[-13.8, -0.9]$; Table 11) is of the same order, so its allocation describes the dip that was observed and does not show that it would recur; over the same steps delivered accuracy rises from 1 to 34 rows (CI $[+15.5, +42.2]$), and in the second harness, with almost no truncation credit (Appendix O), the seed-42 strict curve instead rises ($6.7 \rightarrow 34.2\%$ from step 25 to step 75, Table 42). On the two fresh training seeds, decoded under one engine configuration, the strict dip recurs (-9.1 points, 95% CI -17.7 to -0.9), with a depth that varies with the seed and the engine configuration (Appendix F.5).

Q AN EXTERNAL SCIENCE RECIPE WITH WELL-FORMATTED TARGETS

Four endpoints of a science-domain distillation recipe, whose targets carry boxed answers (16,526 rows that our filtering pipeline selected from the science split of NVIDIA's Llama-Nemotron post-training data (Bercovich et al., 2025), responses by DeepSeek-R1 and Qwen-2.5-32B-Instruct as released; Appendix B), were evaluated at their final checkpoints under the identical protocol (pre-declared exploratory). The four training runs differ only in the model seed and share their data order. Strict and delivered-answer accuracy agree to within $0.13\,\mathrm{pp}$ in 13 of the 16 cells, and within $0.9\,\mathrm{pp}$ in the three NATIVE cells where a few responses ended inside their thought, because the recipe's format is stable (strict parse rates 90–100%, budget hits $\leq 8\%$). Its PRECLOSED$-$NATIVE AIME contrasts across the four endpoints are $+0.9$, -15.5, -7.8 and $-6.9\,\mathrm{pp}$ strict and $+0.9$, -14.7, -6.9 and $-6.9\,\mathrm{pp}$ committed (in two endpoints the strict rule also credits one NATIVE response that ended inside its thought), and its GPQA contrasts are -0.5 to $+2.0\,\mathrm{pp}$ under both scorers. Contract effects on hard mathematics that are not format artifacts therefore appear once format is held constant, in three of the four endpoints under either scorer, and the Gemini effect of §A.3 was not one of them. These per-endpoint AIME contrasts carry no intervals here, and a re-run of the base model under another engine configuration alone changed the sign of its AIME contract contrast (Appendix P), so whether the endpoints, which share their data order, differ from one another is not resolved.

R EXPLORATORY TRAINING EVALUATIONS ON GEMMA-4-31B

Fifteen earlier exploratory evaluation runs on Gemma-4-31B-it were re-scored: the base instruct model and checkpoints distilled from a Gemini-family recipe and from a gpt-oss recipe, the latter from two training runs of the same recipe with the same seed (run 1 read at step 1000, run 2 at four checkpoints). Table 46 lists all fifteen. Seven runs re-decode a checkpoint with a second set of four decoding seeds ("fresh decoding"); items and token budget are those of the Qwen bed. The arms are *bare assistant*, whose prompt ends at the model turn so that the model opens its own thought (it did on all 13,560 scored rows, although as run the AIME system turn omits the thinking-mode token that the GPQA one carries), and PRECLOSED. The judge prompt is the Qwen bed's, but the span read is not. Their strict scores use the rule of the original runs, which reads the text after the last thought closer that does not end a well-formed thought block (the whole output if there is none), so a bare-assistant response is read whole, closed thought included; the judge read the last

3,000 characters of that same span, and its answers were normalised as in the original runs (the first integer or letter; the Qwen bed's rule, which requires the whole answer to be one, differs on 10 of its 40,384 judged rows, all AIME answers that are not integers). In the Qwen bed both read a closed response only after its closer.

On GPQA the gpt-oss-recipe checkpoints give strict contrasts of -6 to -16 pp that are -4 to $+3$ pp committed, and 11–47% of their rows commit without a strict parse, more so under PRECLOSED. Part of this comes from the span: on 87–148 of the 788 bare-assistant GPQA rows per run (11.0–18.8%) the only strict parse is a box inside the closed thought, a parse no PRECLOSED row gets. Scoring strict on the text after the last thought closer, as in the Qwen bed, moves these ten strict contrasts to -3.6 to $+5.2$ pp, so most of their strict contract contrast comes from boxes inside closed thoughts. We then re-judged every row on the Qwen bed's span as well (exploratory, post hoc; same judge, prompt and decoding; on the 13,385 re-judged confirmatory rows whose span does not change, it returns its original answer on 13,383). Their committed contrasts become -2.7 to $+3.9$ pp (CIs above zero in three runs), so under the Qwen bed's reading the contract artifact of these ten runs is -1.3 to $+3.9$ pp, its CI excluding zero in one run (Table 47). The format artifact does not vanish; it becomes nearly symmetric: in every GPQA cell of these runs 19–50% of rows commit an answer without a strict parse, and strict accuracy lies 15–38 points below delivered accuracy in both arms. In the other five runs the strict parses under the two spans differ on at most one row and give identical strict contrasts, and on AIME they never differ.

The base instruct model, by contrast, pays a real price for pre-closing on GPQA: -8.2 and -7.0 pp strict, -6.6 and -7.0 pp committed (CIs exclude zero) under the two sets of decoding seeds. So do the three Gemini-recipe runs, -4.4 to -6.7 pp committed with CIs excluding zero, which the strict grader *understates* at -1.0 to -3.6 pp: the artifact can run in either direction. These five runs, and all fifteen on AIME, give identical contrasts under the Qwen bed's reading, where the Gemini-recipe artifact is $+3.2$ to $+3.4$ pp with CIs excluding zero in all three runs. On AIME strict and committed contrasts agree within 3.5 pp in all fifteen runs. Two qualifications: the gpt-oss recipe's targets *do* box, so training-target format is one source of format drift but not the only one; and the runs were not pre-registered.

Table 46: Gemma-4-31B, all fifteen runs: GPQA PRECLOSED minus bare-assistant accuracy (pp; brackets: unadjusted item-cluster 95% CI) under the original strict rule and delivered scoring, both of which read bare-assistant responses on the original span, closed thought included (see text), and the share of rows committed without a strict parse (bare / PRECLOSED). Run 1 and run 2 are two training runs of the gpt-oss recipe with the same seed; "fresh decoding" re-decodes the same checkpoint with a second set of four decoding seeds. Bare assistant: the prompt ends at the model turn and the model opens its own thought; PRECLOSED: the prompt appends an empty thought; strict: last `\boxed{}` answer on the span stated in the text; committed: the final answer a gold-blind judge reads from the same span; unparsed-committed: rows committing to an answer the strict rule does not parse; Gemini recipe, gpt-oss recipe: distillation recipes named for their teachers; step: optimizer step; pp: percentage points. Base instruct: the untouched Gemma-4-31B-it.

	strict (original rule)	committed	unparsed-committed
run	estimate [95% CI]	estimate [95% CI]	bare / PRECLOSED
base instruct	-8.25 [-11.42, -5.33]	-6.60 [-9.14, -4.19]	6.1 / 9.3
base instruct, fresh decoding	-6.98 [-10.79, -3.43]	-6.98 [-10.15, -4.06]	6.1 / 7.4
Gemini recipe, step 2000	-1.02 [-4.95, $+2.79$]	-4.44 [-7.36, -1.65]	15.9 / 13.3
Gemini recipe, step 2941	-3.55 [-7.36, $+0.25$]	-6.73 [-10.15, -3.43]	11.5 / 9.5
Gemini recipe, step 2941, fresh decoding	-1.78 [-5.58, $+1.90$]	-5.08 [-8.63, -1.78]	12.8 / 10.4
gpt-oss run 1, step 1000	-10.41 [-14.85, -5.96]	$+1.40$ [-2.03, $+4.70$]	26.1 / 42.1
gpt-oss run 1, step 1000, fresh decoding	-10.66 [-15.10, -6.35]	$+3.05$ [-0.38, $+6.47$]	24.4 / 42.9
gpt-oss run 2, step 250	-6.35 [-10.53, -2.16]	-1.02 [-3.68, $+1.65$]	10.9 / 18.5
gpt-oss run 2, step 500	-14.72 [-19.16, -10.41]	-1.02 [-3.81, $+1.90$]	19.9 / 37.7
gpt-oss run 2, step 1000	-12.18 [-16.62, -7.74]	-2.92 [-6.60, $+0.63$]	30.8 / 44.3
gpt-oss run 2, step 2600	-15.99 [-20.56, -11.68]	-2.79 [-5.46, 0.00]	11.7 / 28.8
gpt-oss run 2, step 250, fresh decoding	-9.26 [-12.94, -5.58]	-0.38 [-3.68, $+2.92$]	11.4 / 23.1
gpt-oss run 2, step 500, fresh decoding	-14.72 [-19.42, -10.03]	-3.55 [-6.98, -0.13]	21.1 / 37.3
gpt-oss run 2, step 1000, fresh decoding	-9.01 [-13.71, -4.44]	$+0.76$ [-2.54, $+4.19$]	33.6 / 47.5
gpt-oss run 2, step 2600, fresh decoding	-15.99 [-20.81, -11.29]	-2.92 [-6.35, $+0.38$]	13.5 / 29.9

Table 47: Gemma-4-31B, all fifteen runs, under the Qwen bed's reading (strict and committed both read the text after the last thought closer; exploratory, post hoc): GPQA PRECLOSED minus bare-assistant strict and delivered accuracy (pp), their difference (the contract artifact; brackets: unadjusted item-cluster 95% CI), and strict minus delivered accuracy of each arm (bare / PRECLOSED). Bare: the bare-assistant prompt, in which the model opens its own thought; PRECLOSED: the prompt appends an empty thought; strict: last `\boxed{}` answer after the last thought closer; committed: the final answer a gold-blind judge reads there; Gemini recipe, gpt-oss recipe: distillation recipes named for their teachers; step: optimizer step; pp: percentage points; CI: item-cluster bootstrap 95% interval. Base instruct: the untouched Gemma-4-31B-it; fresh decoding: the same checkpoint re-decoded with a second set of four decoding seeds.

	PRECLOSED − bare		artifact	strict − committed
run	strict	committed	estimate [95% CI]	bare / PRECLOSED
base instruct	−8.25	−6.60	−1.65 [−3.55, +0.25]	−4.44 / −6.09
base instruct, fresh decoding	−6.98	−6.98	0.00 [−2.03, +2.03]	−4.19 / −4.19
Gemini recipe, step 2000	−1.02	−4.44	+3.43 [+0.25, +6.60]	−12.44 / −9.01
Gemini recipe, step 2941	−3.55	−6.73	+3.17 [+0.76, +5.71]	−9.64 / −6.47
Gemini recipe, step 2941, fresh decoding	−1.78	−5.08	+3.30 [+0.51, +6.22]	−10.66 / −7.36
gpt-oss run 1, step 1000	+4.44	+3.55	+0.89 [−3.05, +4.82]	−34.01 / −33.12
gpt-oss run 1, step 1000, fresh decoding	+2.66	+3.93	−1.27 [−5.46, +2.66]	−33.12 / −34.39
gpt-oss run 2, step 250	+3.43	−0.51	+3.93 [+0.13, +7.74]	−19.29 / −15.36
gpt-oss run 2, step 500	+0.63	−0.38	+1.02 [−3.05, +5.08]	−30.46 / −29.44
gpt-oss run 2, step 1000	+3.68	+0.13	+3.55 [−0.63, +7.74]	−37.56 / −34.01
gpt-oss run 2, step 2600	−3.55	−2.66	−0.89 [−4.57, +2.79]	−22.21 / −23.10
gpt-oss run 2, step 250, fresh decoding	+0.63	−0.13	+0.76 [−2.54, +4.19]	−19.29 / −18.53
gpt-oss run 2, step 500, fresh decoding	−0.89	−2.54	+1.65 [−2.66, +5.96]	−30.84 / −29.19
gpt-oss run 2, step 1000, fresh decoding	+5.20	+3.68	+1.52 [−2.79, +5.84]	−38.07 / −36.55
gpt-oss run 2, step 2600, fresh decoding	−1.65	−1.90	+0.25 [−3.68, +4.31]	−23.86 / −23.60

Part V Related work

S Positioning relative to prior work

Table 48 places the closest lines of work against six properties. Extraction fragility, format bias, the no-thinking comparison and the link between termination and measured accuracy in long-CoT SFT (Kopiczko et al., 2026) are established, and Long et al. (2025) and Yu et al. (2025) offer remedies (format-robust fine-tuning, a trained extractor) that we do not. What this paper adds is the dependence of the artifact on training state, the exact identity with a two-sided bound set by two judge-free counts and widened on each side by the parser–judge disagreement term δ, the association of the mechanism with the final-answer act of the training targets (correlational across the two ecological recipes; causal for answer format in a box-removal twin on Qwen3-4B-Base, two training seeds, Appendix N), the demonstration that one grader both manufactures and masks effects along the same trajectory, and prospective tests across items and models. On new items with the same checkpoints, six of nine predictions pass: the judge-free bound fails by one parser–judge disagreement row, one termination cell misses its threshold, and the gpt-oss recipe, predicted to show no artifact, has a nested-box habit (Appendix J). R1-Qwen training reproduces the collapse but not the registered directional predictions (Appendix M.1); R1-Llama reproduces both halves (Appendix M.2); removing only the boxes reproduces the late error in Qwen3-4B-Base (Appendix N); on twelve public models, all three predictions hold in every eligible cell (Appendix L). Exploratory Gemma evaluations extend backbone coverage, without claiming a matched training design (Appendix R).

S.1 Termination in long-CoT SFT

Kopiczko et al. (2026) are closest in setting: long-CoT SFT from base models, AIME and GPQA with the answer requested in `\boxed{}`, and termination reported beside accuracy. They compare separately trained runs that spend a fixed number of updates on more epochs over fewer samples, find that termination (ending with an end-of-sequence token rather than at the token budget) rises with repetition together with accuracy, and attribute the link to unterminated outputs producing no final answer; in their termination-conditioned GPQA table no unterminated generation is counted correct. They decode at each model's recommended sampling parameters; at those of our base model, the

Table 48: Positioning. Object: *deployed* models under input-side variation, SFT *checkpoints*, separately trained *runs*, *RL* training, or pretraining (*PT*) checkpoints. States: separates budget-hit, unclosed and closed rows. Bound: an exact decomposition with a two-sided bound on the extraction artifact, set by two judge-free counts and widened on each side by the parser–judge disagreement term. Targets: mechanism traced to the training targets. Both: shows the grader manufacturing *and* masking effects. ✓ yes, ○ partly, – no. Prospective: an out-of-sample test of the paper's own claims; ours include registered tests on new items (six of nine predictions passing, the judge-free bound failing by one row), R1-Qwen training (collapse reproduced, directional predictions failing), R1-Llama training (both halves reproduced), a box-removal twin on Qwen3-4B-Base, and twelve public models (all three predictions holding) (✓); targets, this paper: ✓ for answer format only, via the twin; xFinder evaluates its extractor on held-out generalization sets (○); Long et al. (2025) also propose a training-side mitigation. depl., ckpt.: deployed models, checkpoints; artifact: strict minus delivered accuracy; judge-free counts: the shares of rows with a strict parse that hit the budget or ended unclosed, and of closed rows without one. MCQ, MC: multiple-choice (questions); OOD: out-of-distribution; RL: reinforcement learning; SFT: supervised fine-tuning; regex: regular-expression extraction.

work	main finding	object	states	bound	targets	both	prospective
Yu et al. (2025)	regex extraction mis-reads; trained extractor	depl.	–	–	–	–	○
Molfese et al. (2025)	MCQ extraction choices change conclusions	depl.	–	–	–	–	–
Hua et al. (2025)	prompt sensitivity vanishes under a judge	depl.	–	–	–	○	–
Jo et al. (2025)	extraction rules reorder reasoning models	depl.	○	–	–	–	–
Hochlehnert et al. (2025)	reasoning scores sensitive to decoding and budget	depl.	○	–	–	–	–
Huang et al. (2025)	verifiers miss answers or are hacked in RL	RL	–	–	–	○	–
Long et al. (2025)	format adherence vs. performance	depl.	–	–	–	–	–
Chandak et al. (2025)	MCQ shortcuts vs. answer matching	depl.	–	–	–	–	–
Ma et al. (2025)	no-thinking competitive under budgets	depl.	○	–	–	–	–
Sun et al. (2025); Jurayj et al. (2025)	budgets change rankings / answer rate	depl.	○	–	–	–	–
Carlon et al. (2026)	report completion and stopped-run answers apart	depl.	○	–	–	–	–
Twist et al. (2026)	trace validity vs. correctness in fine-tuning	ckpt.	○	–	○	–	–
Kopiczko et al. (2026)	termination tracks accuracy as epochs grow	runs	○	–	○	–	–
Ren et al. (2026)	dip-and-recovery with length surge	ckpt.	○	–	–	–	–
Jin et al. (2025)	early OOD peak, later decline	ckpt.	–	–	–	–	–
Luo et al. (2025)	small-model valley with a length surge	runs	–	–	–	–	–
Yang et al. (2025b)	SFT gains mostly format alignment	runs	–	–	○	–	–
Guha et al. (2025); Yeo et al. (2025)	recipe rankings by strict score	ckpt.	–	–	○	–	–
Gu et al. (2025); Wiegreffe et al. (2025)	valid MC format depends on training stage	PT	–	–	–	–	–
Heineman et al. (2025)	checkpoint noise, reduced by averaging	PT	–	–	–	–	–
this paper	strict confounds answer protocol with correctness	ckpt., depl.	✓	✓	✓	✓	✓

Gemini student's step-25 collapse persists but its truncation credit nearly vanishes, while the gpt-oss student's does not; at step 200 the Gemini student still loops at those parameters and its credit returns (Appendix J). We agree that termination drives measured accuracy early in long-CoT SFT, and differ in the object and in what is graded: we follow the checkpoints of one run and compare a strict grader with the delivered answer on every row. Early on the grader credits boxed guesses inside budget-hit and unclosed thoughts, later it misses prose answers in closed rows, so its error changes sign along the run. An output that stops inside an unclosed thought counts as terminated under their criterion; on GPQA that state carries part of our early strict credit (§6). Liu et al. (2025a) likewise find that long-CoT SFT leaves a model that sometimes fails to stop, and gain accuracy by closing its thought at a length threshold, as budget forcing does (Muennighoff et al., 2025); our PRECLOSED arm is the prompt-side analogue, closing the thought before generation. The early failure we observe is non-termination, not the redundant but finished reasoning called overthinking (Chen et al., 2025b). Under greedy decoding our step-25 students loop where the base model almost never does (Appendix J), in line with degeneration under likelihood-maximising decoding (Holtzman et al., 2020) and with distilled students that loop where their teachers rarely do (Pipis et al., 2026); our sampled budget-hit rows are mostly non-repetitive (§6). Luo et al. (2025) report a long-CoT SFT valley for small models: with little data, accuracy falls well below the base model's while responses lengthen sharply under a fixed generation cap, which they attribute to error accumulation; they report neither termination nor truncation rates. Whether part of such a valley is the grader effect measured here is an open question that we have not tested on their models. Valleys also occur where code is graded by execution (He et al., 2025), so a valley need not be a grading artifact.

S.2 VERIFIER ERROR

Huang et al. (2025) show that rule-based verifiers miss correct answers written in other forms, more often for stronger long-CoT models, and Xu et al. (2025) find such false negatives common among model responses in RL data; both study them as a cost to RL. Huang et al. (2025) also follow training: at each RL checkpoint they compare the training reward with an oracle's and find that the two diverge for a fine-tuned model-based verifier that the policy learns to exploit, while rule-based verifiers stay stable. Their object is the net gap between a training reward and an oracle; ours is an evaluation-time strict grader, whose error we decompose across SFT checkpoints into early credit for guesses and late misses of delivered answers, which have opposite signs. A symbolic-equivalence checker, the usual rule-based remedy for format mismatches, can repair our format term but not the state terms (Appendix H). Hochlehnert et al. (2025) show that reported reasoning accuracy is sensitive to decoding, seeds, prompt format and the generation budget, and recommend robust answer extraction; our states make the truncation and extraction terms explicit and show that they move with training. Lee et al. (2026) correct LLM-judge scores for imperfect sensitivity and specificity estimated from human labels; we instead validate our judge four ways and bracket the artifact with judge-free counts (Appendix E).

S.3 FORMAT AND TRAINING STATE

Yang et al. (2025b) attribute most SFT gains in mathematics to format alignment rather than better reasoning, in the spirit of the superficial alignment hypothesis (Zhou et al., 2023). Our gold-box-appending retrain is a controlled instance: appending a boxed gold answer taught the student to box, not to solve (§5). Under a strict grader, part of an apparent format gain may be grader credit, and a later change of answer format can cost score, as the prose answers of the Gemini students do; the two analyses are complementary. The closest precedent for an answer protocol whose validity depends on training stage comes from pretraining: OLMES (Gu et al., 2025) evaluates multiple-choice tasks under both cloze and lettered formulations because models acquire the lettered format only partway through pretraining, and Wiegreffe et al. (2025) locate when that skill appears. There the remedy is to report the better formulation; in generative long-CoT SFT one fixed strict grader errs in opposite directions at different checkpoints, so we score the delivered answer and bound the artifact instead. Heineman et al. (2025) treat checkpoint-to-checkpoint variation in pretraining evaluations as noise, to be reduced by averaging checkpoints or by better metrics; the grader component isolated here is systematic and moves with training state, so averaging nearby checkpoints would largely leave it in place. More generally, a choice of metric can create an apparent phenomenon (Schaeffer et al., 2023), and one answer can be written in many surface forms, the key problem of Biderman et al. (2024); our late misses are an instance of the latter whose size changes with training.

S.4 STATE-WISE EVALUATION

Two recent works also split outputs by the state of the reasoning. Twist et al. (2026) separate answer correctness from the validity of the reasoning trace (valid, empty, missing or truncated, where truncated covers both an unclosed delimiter and the generation limit) across the checkpoints of reasoning models fine-tuned on data without traces; Carlon et al. (2026), studying the reasoning interfaces of released models, ask evaluations to report correct completion before a deadline, the answer read from a stopped run, differences among unfinished runs and the probability of the correct answer separately. Our termination states refine the first split (budget-hit versus unclosed), and our first recommendation (§8) is close in spirit to the second. Neither work measures a strict grader's error against the delivered answer, or how that error changes with training.

S.5 STATISTICS AND VALIDITY

Small benchmarks need care with uncertainty (Miller, 2024; Bowyer et al., 2025; Madaan et al., 2024; Zhang et al., 2026). Two fresh training seeds cannot estimate between-seed variance, so we treat training seeds as fixed replicates, report every seed and give paired artifact sizes. Preregistration (van Miltenburg et al., 2021) guards against forking paths, not against mis-specified measurement or weak construct validity (Bean et al., 2025).