INTERNALIZING THE HALT VECTOR: FULL-VECTOR AND ON-AXIS RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a reasoning model learn earlier stopping by matching only the projection of a steered activation, rather than reconstructing the full vector? We compare full-vector and on-axis reconstruction on DeepSeek-R1-Distill-Qwen-7B, separating reasoning length, mathematical answer correctness and compliance with the requested answer format. An outcome-informed follow-up to an earlier held-out comparison, with development-only selection, evaluated three shortening targets on 2,000 held-out problems. One target established matched shortening within one percentage point: full-vector and on-axis reconstruction reduced thinking by 23.56% and 23.39%, respectively. Their accuracy was equivalent within the prespecified 2.6-point margin under Math-Verify, while a boxed-answer grader favored full-vector reconstruction by 2.80 points. The on-axis model produced more closed, unboxed answers despite an explicit boxing instruction. A post-hoc, model-assessed audit supports a predominantly formatting-based explanation of this discrepancy, subject to assessor uncertainty and parsing sensitivity. Matching remained unresolved at the other two targets. These results do not establish that full-vector reconstruction is necessary for internalizing halting, nor that on-axis training is free of instruction-following costs. They show why comparisons of internalization objectives should distinguish achieved shortening, answer correctness and format compliance. The evidence is conditional on selected checkpoints from one model and does not establish training-seed robustness.

1 INTRODUCTION

Reasoning models often reach a correct answer well before they stop reasoning, and the tokens that follow cost compute without changing the answer (Chen et al., 2024; Sui et al., 2025). Stopping earlier has a price, because further reasoning sometimes recovers an answer that an early stop would lose. One way to install earlier stopping without an inference-time hook is to train the model's early layers so that their activations match a target built from a halt steering direction. Jayabahu & Adeleke (2026) reported this for one model, with the full-vector objective of LoRRA (Zou et al., 2023). Added during generation, the halt direction acts as a causal stopping knob with an accuracy cost (Section 4).

Why objective comparisons need matched shortening. Which activation target should training use? Full-vector reconstruction fits the whole steered activation. On-axis reconstruction penalizes only the error in the displacement's projection onto the halt direction and leaves the other coordinates free (Equation 1). Historical scalar-target arms (hinge objectives) failed at the tested settings, but those comparisons used configurations selected for reconstruction. A fairer comparison meets a second problem: a recipe's settings determine how much it shortens, and shortening trades against accuracy. An accuracy difference between objectives is therefore hard to interpret unless held-out shortening is matched, or the shortening difference is itself reported.

Three held-out comparisons. We report three prospectively frozen comparisons on DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), each selecting recipes on development problems only and testing them on at least 2,000 held-out problems under two graders (Section 5). Every recipe evaluated on held-out problems shortened thinking, by 14.34% to 29.87% of the base model's think tokens, and increased thinking-block closure. In Study 1 (equal selection), no scalar-family candidate

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.

was development-feasible, and the family's on-axis reconstruction fallback shortened more (full-vector minus on-axis cut -3.29 points, below the 5-point minimum of interest), with an accuracy tradeoff under a boxed-answer (legacy) grader; Study 1's early-era project exposure was not certified (Section 5). Study 2 searched a coarse menu for a pair with matched development shortening and found none; its diagnostic finalists differed in held-out cut by $+4.26$ points. Study 3, an outcome-informed follow-up designed after Study 2's development outcomes were known, and frozen only after Study 2's held-out results were known, used 23 values of α per family and three development targets for the cut (12.5, 17.5 and 22.5%). Held-out matching within the 1-point tolerance was established only at 22.5 and is unresolved at 12.5 and 17.5.

The accuracy comparison depends on the grader. At 22.5 the full-vector minus on-axis accuracy difference is -0.04 points under Math-Verify, within the prespecified 2.6-point margin, and $+2.80$ points under the legacy grader; that direction is resolved, but its interval does not settle whether the difference exceeds the 2.6-point threshold of interest. We choose neither grader. Descriptively and post hoc, at 17.5 and 22.5 the grader gap is about the size of the on-axis arm's excess of closed answers that omit the box the prompt requires, which Math-Verify credits and the legacy grader rejects; at 12.5 both are small, and a post-hoc, model-assessed audit supports a predominantly formatting-based reading at 22.5, subject to assessor uncertainty and parsing sensitivity (Section 6.4). The evidence therefore shows neither that the target must constrain more than its halt-direction projection nor that constraining only that projection is free of boxed-answer cost.

Contributions. Jayabahu & Adeleke (2026) reported the halt direction, its hook-free internalization, and a fixed-recipe comparison on one model in which full-vector reconstruction outperformed a hinge-type scalar target, with an off-axis account that we do not rely on. Beyond that work, we contribute:

- **A controlled test of the objective comparison.** On-axis reconstruction, which shares the full-vector target and differs only in the error it penalizes, is compared with full-vector reconstruction under development-only selection and prospectively frozen held-out evaluation, with every null and unresolved result reported (Sections 5 and 6).
- **Correctness and format, separated.** At the matched 22.5 target, accuracy is within the margin under Math-Verify but resolved toward full-vector reconstruction under a boxed-answer grader. A post-hoc, model-assessed audit supports a predominantly formatting-based reading for the audited held-out outputs; separately, closed-output box omission is more frequent for on-axis recipes at all 23 same-α pairs of the development menu (same α is not matched shortening), and the grader split and unboxed excess recur in same-seed and new-seed refits of the 22.5 pair, whose single draw leaves matching unresolved, as it does for the original checkpoints (Section 6.4).
- **A protocol lesson.** Comparisons of length-reduction objectives should match or report achieved shortening and use more than one grader; certifying a 1-pp match at 2,000 problems operates near the test's resolution (Section 6.3).

2 Related work

Steering reasoning and stopping. Representation engineering reads and intervenes on activation directions (Zou et al., 2023), including inference-time intervention (Li et al., 2023), activation addition (Turner et al., 2023; Rimsky et al., 2024) and refusal steering (Arditi et al., 2024). For reasoning models, Sheng et al. (2025) control length through a direction acting on the `</think>` logit, and Venhoff et al. (2025) steer reasoning behaviors in the R1-Distill family; related work studies functional axes recruited by reinforcement learning (Han et al., 2026) and steerable verbal uncertainty (Ji et al., 2025). We use the value axis of Jiang et al. (2026) as a control direction.

Training toward steered activations. We adopt the full-vector L2 reconstruction objective of Low-Rank Representation Adaptation (LoRRA) (Zou et al., 2023), which targets base activations plus a scaled steering vector, together with its separation of edited layers from the targeted layer and its merged, hook-free adapter. CASAL also regresses full steered vectors (Yang et al., 2025b), Ackerman (2024) uses cosine alignment, and static weight edits are another route (Arditi et al., 2024; Sun et al., 2025). We propose no new objective; we ask how full-vector and on-axis reconstruction compare once the shortening each produces is controlled (Appendix L).

Efficient reasoning. Approaches to overthinking (Chen et al., 2024; Sui et al., 2025) include length-penalty reinforcement learning (Arora & Zanette, 2025; Luo et al., 2025; Liu et al., 2025), controllable-length training (Aggarwal & Welleck, 2025), test-time budget control (Muennighoff et al., 2025), decoding-time early exit (Yang et al., 2025a) and supervised compression (Xu et al., 2025). We make no efficiency-frontier claim: DLER (Liu et al., 2025), a length-penalty reinforcement-learning model from the same base, compresses roughly three times as much on the external benchmarks of Appendix J.

Evaluation and prospective design. Reported reasoning gains are sensitive to evaluation details such as sampling seeds and decoding settings (Hochlehnert et al., 2025); we find that the answer grader is one such detail for objective comparisons. Fixing selection rules, tolerances and stopping points before outcomes exist limits undisclosed analytic flexibility (Simmons et al., 2011; Nosek et al., 2018). Training-seed variation, a known source of variance in benchmark comparisons (Bouthillier et al., 2021), is outside our intervals.

3 SETUP

Model, direction and steering. All three studies use `DeepSeek-R1-Distill-Qwen-7B` (Guo et al., 2025) and a halt direction $\hat{u}$ built as in Jayabahu & Adeleke (2026) at layer 18 (Study 1 also tried layer 24; Table 3), a difference of mean activations between think-token positions after and before the probability of the gold answer's first token settles above 0.9 in correctly answered rollouts (Appendix I). Steering adds $a\,\hat{u}$ to the layer-18 residual stream, with $a = (\alpha/100)\|\bar{h}\|$ and $\|\bar{h}\|$ the mean activation norm. Internalization removes the hook by training the model's own layer-18 activations toward a steered target.

Writer, reader and training data. As in Jayabahu & Adeleke (2026), layers 0 to 18 form a *writer* that receives low-rank adapters (Hu et al., 2022), and layers 19 to 27 a frozen *reader*. Every fit uses the same teacher traces for 24 training problems (in Study 1, whose step counts we report, 42 of 48 rows covering 23 problems passed the format check), a language-model loss and one auxiliary target loss, scheduled by the forced-answer point B_{end}, the earliest reasoning prefix whose forced, greedy answer is already correct; B_{end} needs the gold answer, so it is a training label only (Appendix A).

Target objectives. For think token t, let $s(t) \in [0, 1]$ be a schedule weight that reaches 1 at B_{end}, $h(t)$ the adapted layer-18 activation and $h_{\text{base}}(t)$ the same activation with the adapter disabled. The two objectives share one target and differ only in which part of the error they penalize:

$$h^*(t) = h_{\text{base}}(t) + s(t)\,a\,\hat{u}, \quad d(t) = h(t) - h^*(t),$$
$$\mathcal{L}_{\text{R}} = \Big\langle \frac{\|d(t)\|^2}{a^2+\epsilon} \Big\rangle_t, \qquad \mathcal{L}_{\text{S}} = \Big\langle \frac{\big(d(t)^\top \hat{u}\big)^2}{a^2+\epsilon} \Big\rangle_t, \tag{1}$$

where $\langle\cdot\rangle_t$ averages over think tokens and $\epsilon = 10^{-8}$. *Full-vector reconstruction* ($\mathcal{L}_{\text{R}}$, arm R) regresses the whole activation onto the steered one, which pins the off-axis components to their base values; the objective is LoRRA's (Zou et al., 2023), and we adopt it. *On-axis reconstruction* ($\mathcal{L}_{\text{S}}$, arm S) penalizes only the error along $\hat{u}$, so only the language-model loss constrains the off-axis components. In Studies 2 and 3, S is on-axis reconstruction and nothing else. In Study 1, S is a family of scalar targets: hinge and ramp targets on the cosine projection $h^\top \hat{u}/\|h\|$, plus on-axis reconstruction as one member (Appendix A). Results for on-axis reconstruction do not speak for the hinge or ramp targets, or the reverse. In Eq. 1, α sets both the target displacement a and the normalization a^2, so varying it is a local recipe search, not a pure change of strength; Studies 2 and 3 vary only α, at one training seed, and Study 1's menus also vary layer, capacity and schedule.

Decoding. Every evaluation uses one user prompt that requires a boxed final answer (Appendix B) and samples with vLLM (Kwon et al., 2023) at temperature 0.7 and top-p 0.95, with at most 16,384 new tokens and native end-of-sequence stopping. Each draw applies one fixed engine seed and four samples to every problem (Appendix B). A completion's think length is the token count of its text before `</think>`, or of the whole completion when it does not close. Capped, unclosed, wrong and unparseable completions stay in every denominator.

Estimators. For arm a and problem i, m_{ai} is the mean, over draws, of the within-draw median of the four think lengths. The cut of arm a, in percent of base think tokens, and the compression contrast are

$$C_a = 100\Big(1 - \frac{\sum_i m_{ai}}{\sum_i m_{\mathrm{B}i}}\Big), \qquad D = C_{\mathrm{R}} - C_{\mathrm{S}}.$$

Accuracy and closure are averaged over draws and samples within each problem, then equally over problems. Contrasts such as R−S are differences of these rates in percentage points (pp).

Graders and closure. Math-Verify 0.8.0 (Kydlíček, 2025) parses the full completion, reasoning included, and needs no box. The boxed-answer grader (called *legacy* below) requires a box: it normalizes the last `\boxed{}` answer and compares it with the gold answer as a string, then as a number. Closure is the literal presence of `</think>`, not end-of-sequence completion or finishing below the cap. We report both and choose neither.

Uncertainty, margins and multiplicity. Intervals come from a paired one-stage bootstrap over problems (Efron, 1979) in which each resampled problem carries every arm, draw and sample (200,000 replicates); they are conditional on the selected checkpoints and exclude training-seed variation. Studies 1 and 2 use two-sided 95% intervals for directions and one-sided 95% bounds for safeguards and tolerances. The accuracy margin of 2.6 pp and the closure margin of 1 pp were fixed before any held-out outcome; they are tolerances, not estimates of noise. Study 3 applies a Bonferroni divisor of 3 over its targets within each endpoint family: one-sided $98\frac{1}{3}\%$ bounds for matching, safeguards and positive cuts, and two-sided $98\frac{1}{3}\%$ intervals for accuracy direction and magnitude; a resolved accuracy direction whose interval lies wholly below 2.6 pp is tagged below the threshold of interest. This is not a simultaneous guarantee across endpoint families, graders or arms, and per-draw summaries are descriptive because all draws share the same problems and bootstrap seed (rules: Appendix B).

4 Internalization shortens reasoning

This section summarizes earlier fixed-recipe evidence, whose estimators and holdouts differ from ours (Appendices I and J). Added during generation, the halt direction changes reasoning length: of the four layers Jayabahu & Adeleke (2026) sweep, layer 18 controls length most strongly ($\rho = -0.82$ between strength and length), in a separate layer-6 sweep a magnitude-matched value-axis control barely moves length ($\rho = +0.09$), and in our runs removing the direction selectively disrupts stopping. It is a stopping actuator with an accuracy cost, not a demonstrated endogenous halting representation, and it does not explain why objectives differ.

A low-rank adapter on layers 0 to 18, trained toward full-vector targets at $\alpha = 25$ on 24 selected problems, removed 20 to 24% of thinking zero-shot on a 100-problem holdout and four external benchmarks, with accuracy changes from -2.0 to $+2.7$ points, no reported intervals and no established separation from all development data (Jayabahu & Adeleke, 2026). In our runs at 1.5B and 14B, each selecting its recipe on its own holdout, the recipe also shortens reasoning, with a resolved AIME25 accuracy cost at both sizes, and on R1-Distill-Llama-8B reconstruction gave a null (Appendix J.4). Scalar-projection (hinge) targets failed at the tested settings, but in fixed-recipe ablations conditional on a reconstruction winner, and they are not on-axis reconstruction (Appendix J.5); on a refusal direction a full-vector and a scalar-projection target both installed the behavior (Appendix K). Choosing a target therefore requires a comparison that controls how much each recipe shortens.

5 Three held-out comparisons

The three studies share the setup of Section 3 and one template (Table 1): each family's candidate recipes were evaluated on a development panel, a rule fixed in advance chose finalists from development results alone, and the finalists were sampled on a disjoint held-out panel under an analysis fixed before held-out generation (rules: Appendix B; execution: Appendix H).

Study 1: equal selection. Study 1 asks which objective shortens held-out reasoning more while passing prespecified quality safeguards, given the same incremental selection opportunity. Each

family received seven candidates with a common seed and fitting inputs: full-vector variants of layer, α, capacity and schedule, and scalar hinge and ramp targets plus one on-axis reconstruction member, S7 (Appendix A). Each family sent its most-compressing development-feasible candidate (Math-Verify accuracy at most 2.6 pp below base, closure not below base, a positive cut), or, with none feasible, a flagged fallback. The primary endpoint is D; each finalist faces six safeguards (accuracy under each grader and closure, against base and against the other finalist), and a resolved contrast with $|D| < 5$ pp is below the prespecified minimum of interest and may not be read as a practically meaningful difference (cells: Appendix B).

Study 2: matched natural shortening. Study 1 confounds the objective with how far each finalist shortens; Study 2 fixes the shortening and compares accuracy. R and S are full-vector and on-axis reconstruction (Equation 1), each trained at $\alpha \in \{5, 10, 15, 20, 25, 35, 50\}$; at each α only the loss differs. A pair was eligible if both candidates were feasible under both graders, both development cuts lay within 2.5 pp of a 25% target and $|D| \leq 1$ pp; with no eligible pair, each family's candidate closest to the target was evaluated held out as a diagnostic only. Held out, matching requires both one-sided bounds on D inside ± 1 pp; only then is each grader's R$-$S accuracy contrast read as favoring R or S, within the 2.6 pp margin (state E), or unresolved (rules: Appendix B).

Study 3: a finer, outcome-informed follow-up. Study 3 repeats Study 2 on a finer menu. It is an outcome-informed follow-up, designed after Study 2's development outcomes were seen and frozen after Study 2's held-out results, including the diagnostic finalists' held-out cuts, were known, and not a rescue: Study 2's cell stands. Each family has 23 values of α, every integer from 5 to 25 plus 35 and 50. The seven values shared with Study 2 reuse its adapters, and all conditions were sampled afresh on a 200-problem development panel made of Study 2's 100 development problems and 100 new ones, which reuses no stored Study 2 rows; some conditions reuse its sampling seed, configuration and order and may reproduce previously seen samples, and regeneration removes neither prior exposure nor knowledge of the candidates. Three development targets, 12.5, 17.5 and 22.5% (windows $[10, 15]$, $[15, 20]$, $[20, 25]$), each select a pair independently by Study 2's rule; a target without an eligible pair would be reported untested. Held-out intervals are Bonferroni-adjusted over the three targets. Targets are development labels, not held-out operating points.

Populations. All panels hold competition problems from one public source (Numina, 2024), under study-specific eligibility standards. Study 1 excluded every problem with a documented prior use available to us and kept fitting, development and test inputs disjoint; earlier uses in this line of work are not documented well enough to rule out, so Study 1's *early-era project exposure was not certified*. This concerns earlier use of problems, not pretraining. Studies 2 and 3 admitted problem groups disjoint from documented project uses under declared checks, not universally unexposed groups, and Study 3's held-out panel consists entirely of groups that Study 2's population rule excluded. The populations are not interchangeable, so we never pool or compare them as one population.

6 Results

6.1 Study 1: equal selection

No scalar candidate was feasible at development, so the scalar finalist is the fallback S7, the on-axis member, flagged `NO-DEV-FEASIBLE`; the full-vector finalist R2 (layer 18, $\alpha = 25$) was feasible. R2 shortens held-out reasoning by 26.58% [25.57, 27.57] and S7 by 29.87% [28.74, 30.97], so $D = -3.29$ pp $[-4.20, -2.37]$. The on-axis fallback shortens more; the difference is resolved but below the 5 pp minimum of interest. The registered cell is `S-TRADEOFF` with `RESOLVED-BELOW-INTEREST`. *Study 1's early-era project exposure was not certified* (Section 5), and that uncertainty remains unresolved.

S7's tradeoff rests on the legacy components. Its legacy accuracy changes by -6.93 pp against base and -6.19 pp against R2, and both one-sided bounds lie beyond the 2.6 pp margin (state F, a resolved failure), the only failed safeguards in the three studies. Under Math-Verify, its change against base (-2.13 pp) is unresolved at that margin, and its deficit against R2 (-1.93 pp) passes it. R2 passes all six safeguards (legacy change against base -0.74 pp, a point estimate, not a bound on possible

Table 1: Design of the three held-out studies. R is full-vector reconstruction; S is a scalar-target menu that includes on-axis reconstruction (Study 1) or on-axis reconstruction alone (Studies 2 and 3). Per family: 7, 7 and 23 candidates; panels are problems × draws × samples. Study 1's early-era project exposure was not certified.

	Study 1	**Study 2**	**Study 3**
Comparison	Equal selection	Matched natural shortening	Finer matched follow-up (outcome-informed)
Development panel	$100 \times 1 \times 4$	$100 \times 1 \times 4$	$200 \times 4 \times 4$
Selection rule	Per family, largest feasible development cut	Both cuts in $25 \pm 2.5\%$, $\lvert D\rvert \leq 1$ pp	Per target T, both cuts in $T \pm 2.5\%$, $\lvert D\rvert \leq 1$ pp
Development outcome	R2 feasible; no scalar candidate feasible (fallback S7)	0 of 49 pairs eligible (diagnostic finalists R5/S2)	3, 6 and 6 of 529 pairs eligible
Held-out panel; rows	$2{,}000 \times 2 \times 4$; 48,000	$2{,}130 \times 2 \times 4$; 51,120	$2{,}000 \times 4 \times 4$; 224,000
Frozen outcome	S-TRADEOFF with RESOLVED-BELOW-INTEREST	NO-DEV-MATCH: DIAGNOSTIC-ONLY	MATCH-UNRESOLVED (12.5, 17.5); GRADER-SPLIT (22.5)

loss); a noninferiority pass does not show equal accuracy. Both finalists raise closure against base, by $+6.15$ and $+6.43$ pp. Per-draw results are in Appendix C.

6.2 Study 2: matched natural shortening

No development pair was eligible (0 of 49): the two shortening restrictions alone empty the set, and only 2 of 7 on-axis candidates were feasible, each from $\alpha = 15$ upward failing legacy retention (Appendix D). The one feasible pair with both cuts in the window, R5 ($\alpha = 25$) and S2 ($\alpha = 10$), differed by $+2.50$ pp, beyond the 1 pp tolerance, so the cell is the permanent `NO-DEV-MATCH:DIAGNOSTIC-ONLY`.

Held out, R5 shortens reasoning by 26.78% [25.78, 27.77] and S2 by 22.52% [21.48, 23.54], so $D = +4.26$ pp $[+3.46, +5.07]$: the development mismatch reproduces held out, and the matching component fails. Both finalists pass all six safeguards (a pass does not show preserved accuracy), with small accuracy changes against base: -0.60 and $+0.38$ pp under Math-Verify, -0.87 and -1.56 pp under legacy grading, where S2's two-sided interval reaches -2.27 pp; these points are not bounds on possible loss. The graders favor opposite finalists on R$-$S accuracy, but with shortening unmatched this contrast does not rank the objectives.

6.3 Study 3: a finer matched follow-up

Study 3's selection and held-out rules were frozen before any Study 3 outcome existed and allow no post-test reselection, widened tolerance or best-target-only summary, so we report all three targets (development labels, not held-out operating points) with the adjusted intervals of Section 3. We write R_α and S_α for full-vector and on-axis reconstruction at recipe parameter α.

Selection. Every target had development-eligible pairs (3, 6 and 6 of 529), and the frozen tie-break selected $\mathrm{R}_{10}/\mathrm{S}_6$ at 12.5, $\mathrm{R}_{14}/\mathrm{S}_8$ at 17.5 and $\mathrm{R}_{21}/\mathrm{S}_{11}$ at 22.5. At development, 21 of 23 full-vector and 7 of 23 on-axis recipes were feasible (Appendix E).

Held-out shortening and matching. Every selected arm shortened held-out thinking; the absolute-cut endpoint (a cut above zero) passes for all six arms. Observed cuts lie above their development labels by $+0.89$ to $+2.20$ pp, a descriptive deviation (Table 2). Matching within the 1-pp tolerance is established only at 22.5: $D = C_R - C_S$ is $+0.17$ pp, with adjusted one-sided bounds $[-0.52, +0.86]$ inside ± 1. At 12.5 (D $+0.33$, $[-0.41, +1.08]$) and 17.5 (D -0.83, $[-1.61, -0.03]$) matching is neither established nor rejected, and the frozen cell is MATCH-UNRESOLVED at both. This is near the test's resolution (descriptive arithmetic on the adjusted bounds): at 2,000 problems each bound

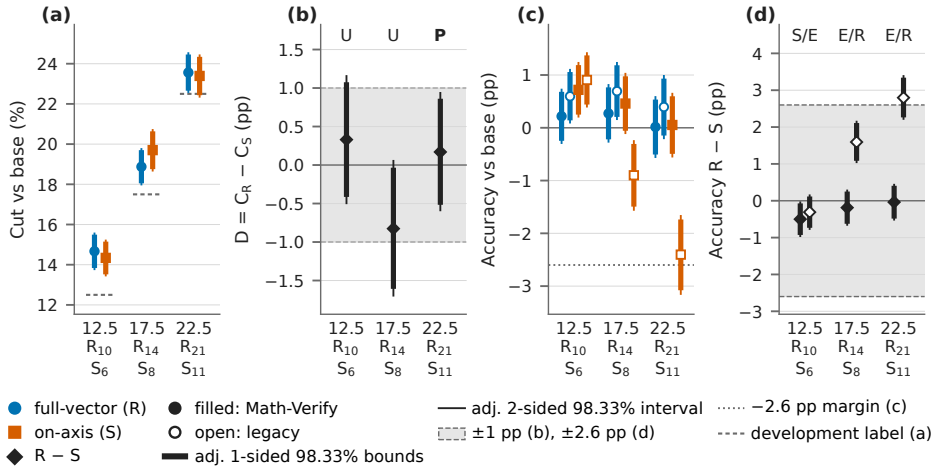

Figure 1: Study 3 held-out results (2,000 problems). The labels 12.5, 17.5 and 22.5 are development targets of one study, not independent replications. (a) Cut against base, full-vector (circles) and on-axis (squares); dashes mark the labels. (b) $D = C_R - C_S$, the ± 1 pp tolerance (shaded) and the match state (P matched, U unresolved). (c) Accuracy change against base, Math-Verify filled, legacy open; dotted line, the -2.6-pp noninferiority boundary. (d) Full-vector minus on-axis accuracy, the ± 2.6 pp margin and the Math-Verify/legacy states of Table 2. Thick bars: adjusted one-sided $98\frac{1}{3}\%$ bounds; thin lines: adjusted two-sided $98\frac{1}{3}\%$ intervals. Points are not bounds on possible loss.

lies 0.69 to 0.79 pp from its point, so matching can be established only when $|D|$ is below about 0.21 to 0.31 pp.

Accuracy by grader. Under Math-Verify (Figure 1, Table 2), the only resolved direction is at 12.5 and favors on-axis reconstruction (-0.50 pp, adjusted two-sided $[-0.98, -0.02]$), below the 2.6-pp threshold of interest; at 17.5 and 22.5 the difference lies within the 2.6-pp margin (-0.19, $[-0.68, +0.30]$, and -0.04, $[-0.54, +0.46]$). Under the legacy grader the difference is within the margin at 12.5 and resolved toward full-vector reconstruction at 17.5 ($+1.59$, $[+1.02, +2.17]$, below the threshold of interest) and at 22.5 ($+2.80$, $[+2.20, +3.41]$), where the interval straddles 2.6 (magnitude unresolved). At the matched target the graders reach different states and the frozen cell is GRADER-SPLIT; no cell assigns an accuracy advantage to either objective at any target. All three targets carry the grader-split and draw-sign-split flags (some contrasts' per-draw points take both signs; Appendix F).

Safeguards and closure. The quality safeguards pass for full-vector reconstruction at 12.5 and 22.5 and for on-axis reconstruction at 12.5 and 17.5. They are unresolved for full-vector reconstruction at 17.5 (closure against on-axis, adjusted lower bound -1.22 against the -1 pp margin) and for on-axis reconstruction at 22.5 (legacy accuracy, adjusted lower bounds -3.08 against base and -3.34 against full-vector reconstruction, below -2.6 pp). No component fails, and a quality pass does not show that accuracy is preserved. Against base, every selected arm raises closure by $+4.60$ to $+7.41$ pp (adjusted lower bounds at least $+4.07$) and changes Math-Verify accuracy by $+0.02$ to $+0.72$ pp (adjusted lower bounds at least -0.51); these points are not bounds on possible loss (Appendix E).

6.4 Grader dependence

Math-Verify, as configured, extracts one answer candidate from anywhere in the completion, reasoning included, without requiring a box (Section 3); the legacy grader credits only a boxed answer. By construction (Appendix G), the legacy-minus-Math-Verify gap in the R−S accuracy difference equals the on-axis arm's excess of outputs credited by Math-Verify alone, net of legacy-only credit: 0.19, 1.78 and 2.83 pp at 12.5, 17.5 and 22.5.

Table 2: Study 3 held-out results per development target (labels, not held-out operating points). Cuts are in percent of base think tokens; $D = C_R - C_S$ has adjusted one-sided bounds and is matched if both lie inside ± 1 pp. Accuracy is full-vector minus on-axis with adjusted two-sided intervals. States: P pass, U unresolved, E within the 2.6-pp margin, R or S resolved toward full-vector or on-axis reconstruction.

Development target label (% cut)	12.5	17.5	22.5
Selected α, full-vector / on-axis	10 / 6	14 / 8	21 / 11
Held-out cut C_R (%)	14.67	18.87	23.56
Held-out cut C_S (%)	14.34	19.70	23.39
$D = C_R - C_S$ (pp)	+0.33	−0.83	+0.17
adjusted one-sided 98.33% bounds	[−0.41, +1.08]	[−1.61, −0.03]	[−0.52, +0.86]
match state (bounds within ± 1 pp)	U	U	P
Math-Verify accuracy, R−S (pp)	−0.50	−0.19	−0.04
adjusted two-sided 98.33%	[−0.98, −0.02]	[−0.68, +0.30]	[−0.54, +0.46]
direction state	S	E	E
Legacy accuracy, R−S (pp)	−0.31	+1.59	+2.80
adjusted two-sided 98.33%	[−0.79, +0.17]	[+1.02, +2.17]	[+2.20, +3.41]
direction state	E	R	R
Quality safeguards, R / S	P / P	U / P	P / U
Frozen cell	MATCH-UNRESOLVED	MATCH-UNRESOLVED	GRADER-SPLIT

This decomposition is post hoc and descriptive. The legacy grader credits fewer outputs in every arm: on the base model, Math-Verify credits 0.789 of outputs and the legacy grader 0.669. Math-Verify credit without a closed `</think>` is rare (39 to 99 outputs per selected arm) and accounts for almost none of the gap (Table 21). We examined one category of Math-Verify-only outputs: completions that close `</think>`, contain no `\boxed` anywhere, and are credited by Math-Verify but not by the legacy grader. The on-axis arm's excess of such outputs is 0.28, 1.86 and 2.82 pp of outputs, about the size of the gap at 17.5 and 22.5 and small at 12.5 (at 22.5, 1,113 against 210 of 32,000 per arm); Study 1 shows a descriptive parallel. A post-hoc, exploratory audit by three language-model assessors, one failing its validity checks (Appendix G), estimates that, among these outputs at 22.5, the on-axis arm has 898 more correct stated final answers than the full-vector arm (95% interval 882 to 904 under primary assumptions that take majority labels as correct and on-axis sampling as the only uncertainty; 369 to 1,004 under the worst-case assessor envelope); outputs are not paired across arms, so this compares two arm totals, not a count within the 903-output difference. This supports a predominantly formatting-based reading, subject to assessor uncertainty and parsing sensitivity; assessor agreement is not accuracy, and there is no human ground truth.

Across the development menu (post hoc, exploratory). On Study 3's development outputs for all 46 recipes (Appendix N), the on-axis recipe omitted the box among closed outputs more often than the full-vector recipe at all 23 same-α pairs, by 0.19 to 6.87 pp, and had more grader disagreement, including the feasible on-axis recipes; the pattern is not specific to the recipes that failed feasibility. Every on-axis recipe with $\alpha \geq 12$ failed development feasibility on legacy accuracy (pooled omission among closed outputs 6.41% against 1.07% for full-vector recipes at the same α), and the association between feasibility failure and grader disagreement is partly definitional because legacy accuracy is a feasibility criterion. Same α does not match shortening, the menu is not a causal dose response, and no development answer was adjudicated, so the audit's correctness assessment does not transfer to these outputs.

Refits of the matched pair (post hoc, descriptive). A study designed after the Study 3 readout refit the 22.5 pair at the original training seed and at one new seed and evaluated each once on the held-out problems (Appendix M). On that draw the original checkpoints, the same-seed refit and the new-seed refit all read within the margin under Math-Verify and resolved toward full-vector reconstruction under the legacy grader (legacy R−S +2.30 to +3.26 pp), and the on-axis excess of closed, unboxed, Math-Verify-only outputs recurred (213 to 248 outputs). Matching reads unresolved in all three, including the original checkpoints: the adjusted one-sided bounds on D span 2.50 to 2.62 pp, wider than the 2-pp band, so none of these evaluations establishes matching. Every point $|D|$

was at most 0.61 pp, but the point D moved by $+0.34$ pp at the same seed and $+1.09$ pp at the new seed, as much as the matching tolerance, so these draws show neither that matching is stable across refits nor that it is lost. Differences combine seed, training nondeterminism and evaluation-path variation (an unchanged adapter did not reproduce its outputs byte for byte). With one new seed, this is fixed-recipe refit evidence, not a replication of the search or evidence of training-seed robustness.

We treat neither grader as correct, and none of these analyses changes a frozen state or cell. Because the prompt requires a boxed answer, these outputs also depart from the instruction; the on-axis excess grows across the three targets (Table 24), so the grader split is also a descriptive difference in answer format. Whether such an answer should count as correct is a grading-policy choice the comparison inherits.

7 Discussion and limitations

What the evidence does and does not support. Jayabahu & Adeleke (2026) argued from a one-model comparison that internalizing a halt direction requires reconstructing the whole steered activation. The evidence here does not establish that requirement: at the one target where matching was established, on-axis reconstruction reached matched held-out shortening with Math-Verify accuracy within the 2.6 pp margin and no failed safeguard, although the legacy difference is resolved toward full-vector reconstruction ($+2.80$ pp, magnitude relative to 2.6 pp unresolved). Nor is on-axis reconstruction shown to be free of cost: Study 1's on-axis arm has the only resolved safeguard failures (both legacy), every on-axis recipe with $\alpha \geq 12$ failed development legacy retention in Study 3, its unboxed answers depart from the prompt's instruction, and at the matched target its legacy accuracy against base is unresolved at the margin (adjusted lower bound -3.08). We report no mechanism: off-axis drift was not measured in the held-out studies, and the attempted selective-restoration study was not validated (Appendix K).

Matching is the difficult step. The recipe parameter moves shortening and accuracy together. On-axis recipes stopped passing the development legacy retention check at moderate α in Studies 2 and 3 (in Study 3 from $\alpha = 12$, the only check they failed up to 25); descriptively they also omit the box more often among closed outputs, although so do feasible on-axis recipes at smaller α (Section 6.4). The region where both objectives are feasible and shorten within 1 pp of each other was therefore narrow: no coarse pair qualified, each fine target had few eligible pairs, and held-out matching was established at one of three targets, near the test's resolution (Section 6.3).

Limitations. Our intervals are across-problem bootstraps conditional on the selected checkpoints, with one training seed per recipe; they exclude training-seed variation. The fixed-recipe refits reproduce the grader pattern at the 22.5 pair but leave matching unresolved on the evaluated draw, including for the original checkpoints; they neither replicate the search nor establish seed robustness, so matching and the R$-$S contrasts are properties of the selected checkpoints, not established properties of the objectives. Matching concerns aggregate thinking length, not which problems are shortened (Appendix E.6), tails, total tokens or wall-clock cost. Earlier recipe searches were not balanced across families, and equal selection equalizes only the incremental search. Study 3's operational amendments during collection (two deadline changes, an overridden launch-scheduling gate whose no-go record is preserved, one replacement per technical failure) changed no pair, target, sample, draw, grader, tolerance or analysis rule (Appendix H). The grader audit uses language-model assessors without human ground truth. The held-out comparisons use one 7B model, one layer, one prompt format, one token cap and mathematical problems (Appendix L); training needs gold-answer stopping labels computed offline; and hook-free deployment establishes neither wall-clock savings nor deployment safety.

8 Conclusion

Training toward halt-derived activation targets shortened held-out reasoning but did not settle which target to use: at the one target with established matching, accuracy was within the margin under Math-Verify while a boxed-answer grader favored full-vector reconstruction, tracking unboxed on-axis answers. Objective comparisons should match or report shortening and separate answer correctness from format compliance.

AI USE STATEMENT

The research question originated with the authors, and the motivating hypothesis, that internalizing a steering direction might require full-vector reconstruction, was drawn from the one-model comparison of Jayabahu & Adeleke (2026); neither was generated by an AI tool. The held-out results reported here do not establish that hypothesis, and the paper reports them as found. An author reviewed each study's design before its held-out evaluation, decided every amendment made during data collection (each is disclosed in Appendix H), and made the scientific judgments the paper rests on, including which results it stands behind and which it reports as unresolved. Generative AI tools were used throughout the execution of the work. We list that use in the two categories the ICLR policy distinguishes.

Required disclosure. Of the twelve categories the policy lists, seven apply. We used generative AI to refine the authors' hypothesis into the prospectively frozen form in which it was tested, to design experiments and give feedback on methodology, to implement methods, to clean and reformat data for the training and evaluation pipelines, and to interpret results. The training and evaluation code, the frozen protocol documents, the job scheduling and verification tooling, and the analysis code behind the reported numbers were written with AI assistance under author direction, and AI agents operated the evaluation runs under rules the authors set. In the post-hoc grader audit, language models served as assessors that labeled whether audited outputs state a correct final answer (Appendix G); we count this use under implementing methods and interpreting results, and those labels are model assessments, not human judgments. Two further categories apply to items co-produced rather than delegated: formulating mathematical claims (the definitions of the reported metrics) and developing the conceptual framework surrounding the central hypothesis. The other five categories do not apply. We did not use a generative AI tool to generate synthetic data: the problem sets are existing public datasets, and the training pools are built from the study model's own rollouts on those problems, which is the system under investigation rather than an authoring aid. The work contains no mathematical proofs, no translation, and no qualitative or thematic data analysis; the assessors' correctness labels are a measurement disclosed above.

Recommended disclosure. We used generative AI to draft and edit the text of this paper, to propose its structure, candidate titles and keywords, to write the figure- and table-generating code and other software, to prepare the code artifact for release, to search for and summarize related work, and to format references. The final title and the ordering of results are the authors' decisions.

Review and verification. We have reviewed all AI-assisted work. Every result number from the three held-out studies is generated by script from the frozen analysis outputs rather than typed, and the primary Study 3 analysis was checked against an independent reimplementation (Appendix H). Historical results summarized from earlier work keep the provenance caveats stated with them (Appendix J). We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects, no personal data and no crowdsourced annotation; all problem sets are public benchmarks. Its object is a training method that changes how long a reasoning model thinks before answering. Shorter reasoning can lower accuracy on some problems, and the accuracy effects we report depend on the grader, so a deployment that relies on internalized halting should check correctness and termination on its own task distribution. The same activation-target training also installs refusal when trained toward a refusal direction (Appendix K), so it can write other behaviors, including safety-relevant ones, into a model's weights without any inference-time hook; such edits should be evaluated for side effects in the same way as other fine-tuning. We report every null and unresolved result of the held-out studies alongside the positive ones.

REPRODUCIBILITY STATEMENT

Held-out studies. For each of the three held-out studies, the candidate menus, selection rule, tolerances and margins, development and held-out panels, draws and seeds, graders, estimators, bootstrap and decision tables were frozen in versioned documents before any held-out generation (Section 5; Appendices A and B). Changes made during collection are listed with dates in Appendix H. All three studies use DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), vLLM sampling at temperature 0.7 and top-p 0.95 with a 16,384-token cap, Math-Verify and a boxed-answer legacy grader, and a paired problem bootstrap. For Study 3: Math-Verify 0.8.0 with SymPy 1.13.1; development on 200 problems with draw seeds 101, 307, 401 and 503; held-out evaluation on 2,000 problems with draw seeds 173, 271, 619 and 827, four samples per problem per draw; bootstrap with NumPy's PCG64 generator seeded 543, 200,000 replicates and linear quantiles. Every result number from these studies is generated by script from the frozen analysis outputs. Validation differs by study: Study 3's primary analysis agrees with an independent reimplementation, while Studies 1 and 2 carry the qualified cross-check and scorer disclosures of Appendices C and D.

Post-hoc analyses. The grader audit (Appendix G), the fixed-recipe refit study (Appendix M), the development formatting analysis (Appendix N) and the per-problem analysis (Appendix E.6) each have a dated specification (Appendix H); the refit study used Study 3's held-out draw seed 173 and training seeds 20260918 and 3049121431.

Historical results. The published recipe (Jayabahu & Adeleke, 2026) fits 24 selected problems; its direction construction, stopping labels and recipe selection rely on a larger offline pipeline whose cost is not reported. Our runs of that recipe at other model sizes are described in Appendix J.

Code and data. An anonymous repository for review, `https://anonymous.4open.science/r/halt-internalization`, and the identical supplementary archive contain the analysis code, frozen protocol configurations, package versions, held-out problem identifiers and aggregate analysis outputs, from which the generated numbers and tables of the three studies and the post-hoc analyses can be regenerated on a CPU (historical results summarized from earlier work are not), together with reference copies of the training, evaluation and grader code used in the studies. They omit trained adapters, the halt direction, teacher traces, row-level generations and scores, and the job-management layer (including the seeding wrapper of Appendix A), so they do not reproduce the GPU runs end to end. The problem sets are public. Row-level generations are retained.

REFERENCES

C. Ackerman. Representation Tuning. *arXiv preprint arXiv:2409.06927*, 2024.

A. Afzal, F. Matthes, G. Chechik, and Y. Ziser. Knowing Before Saying: LLM Representations Encode Information About Chain-of-Thought Success Before Completion. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

P. Aggarwal and S. Welleck. L1: Controlling How Long a Reasoning Model Thinks with Reinforcement Learning. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2503.04697.

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.11717.

D. Arora and A. Zanette. Training Language Models to Reason Efficiently. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.04463.

C. Blank, A. Bhatia, S. Rajamanoharan, A. Conmy, and N. Nanda. Subliminal learning is steering vector distillation. *arXiv preprint arXiv:2606.00995*, 2026.

X. Bouthillier, P. Delaunay, M. Bronzi, et al. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

X. Chen et al. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. *arXiv preprint arXiv:2412.21187*, 2024.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint*, 2018. arXiv:1803.05457.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

C. Fierro and F. Roger. Steering Language Models with Weight Arithmetic. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.05408.

D. Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. arXiv:2501.12948.

A. Q. Han, D. J. Chalmers, and P. Izmailov. How's It Going? Reinforcement Learning in Language Models Recruits a Functional Welfare Axis. *arXiv preprint arXiv:2605.30232*, 2026.

D. Hendrycks et al. Measuring Mathematical Problem Solving with the MATH Dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

A. Hochlehnert, H. Bhatnagar, V. Udandarao, et al. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

E. J. Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

G. Ilharco et al. Editing Models with Task Arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

D. Jayabahu and T. Adeleke. The Halt Vector: Internalizing a Causal Steering Intervention for Efficient Reasoning. In *2nd Workshop on Efficient Reasoning, Conference on Language Modeling (COLM)*, 2026. Non-archival workshop paper. arXiv:2608.28859.

Z. Ji et al. Calibrating Verbal Uncertainty as a Linear Feature to Reduce Hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3769–3793, 2025. arXiv:2503.14477.

N. Jiang, I. Kauvar, and J. Lindsey. The Value Axis: Language Models Encode Whether They're on the Right Track. *arXiv preprint arXiv:2606.17056*, 2026.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

H. Kydlíček. Math-Verify: A robust mathematical expression evaluation system. https://github.com/huggingface/Math-Verify, 2025.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03341.

S.-Y. Liu et al. DLER: Doing Length pEnalty Right, Incentivizing More Intelligence per Token via Reinforcement Learning. *arXiv preprint arXiv:2510.15110*, 2025.

H. Luo et al. O1-Pruner: Length-Harmonizing Fine-Tuning for O1-Like Reasoning Pruning. *arXiv preprint arXiv:2501.12570*, 2025.

N. Muennighoff et al. s1: Simple Test-Time Scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 20275–20321, 2025. arXiv:2501.19393.

B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Numina. NuminaMath-CoT. https://huggingface.co/datasets/AI-MO/NuminaMath-CoT/tree/9d8d210c9f6a36c8f3cd84045668c9b7800ef517, 2024.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2311.12022.

N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. M. Turner. Steering Llama 2 via Contrastive Activation Addition. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

L. Sheng, A. Zhang, Z. Wu, W. Zhao, C. Shen, Y. Zhang, X. Wang, and T.-S. Chua. On Reasoning Strength Planning in Large Reasoning Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.08390.

Z. Shi, Y. Zhu, J. Shi, X. Zhang, L. Wang, and C. Miao. Internalizing llm reasoning via discovery and replay of latent actions. *arXiv preprint arXiv:2602.04925*, 2026.

J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22 (11):1359–1366, 2011.

Y. Sui et al. Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2503.16419.

C.-E. Sun, G. Yan, and T.-W. Weng. ThinkEdit: Interpretable Weight Editing to Mitigate Overly Short Thinking in Reasoning Models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. arXiv:2503.22048.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, and J. Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. In *Findings of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2210.09261.

A. M. Turner et al. Steering Language Models with Activation Engineering. *arXiv preprint arXiv:2308.10248*, 2023.

C. Venhoff, I. Arcuschin, P. Torr, A. Conmy, and N. Nanda. Understanding Reasoning in Thinking Language Models via Steering Vectors. *arXiv preprint arXiv:2506.18167*, 2025.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.01574.

J. Wehner, S. Abdelnabi, D. Tan, D. Krueger, and M. Fritz. Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models. *arXiv preprint arXiv:2502.19649*, 2025.

M. Wortsman et al. Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy without Increasing Inference Time. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2203.05482.

S. Xu, W. Xie, L. Zhao, and P. He. Chain of Draft: Thinking Faster by Writing Less. *arXiv preprint arXiv:2502.18600*, 2025.

P. Yadav et al. TIES-Merging: Resolving Interference When Merging Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.01708.

C. Yang et al. Dynamic Early Exit in Reasoning Models. *arXiv preprint arXiv:2504.15895*, 2025a.

W. Yang, X. Qiu, L. Yu, Y. Zhang, A. Yang, N. Kokhlikyan, N. Cancedda, and D. Garcia-Olano. Hallucination Reduction with CASAL: Contrastive Activation Steering for Amortized Learning. *arXiv preprint arXiv:2510.02324*, 2025b.

A. Zhang et al. Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.05419.

A. Zou et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405v1*, 2023.

A OBJECTIVES, ADAPTERS AND TRAINING

Direction and steering. A precomputed halt direction $\hat{u}$ is fixed for each target layer and shared across fits and studies (its construction is in Appendix I); every held-out finalist uses layer 18, and only Study 1's menu includes layer-24 candidates (Table 3). Steering at layer ℓ adds $a\,\hat{u}$ with $a = (\alpha/100)\|\bar{h}\|$, where $\|\bar{h}\|$ is the mean activation norm. The training targets below reuse the same a, so a recipe's α is in the same units as inference-time steering.

Schedules. Let k be a think token's rank and $b = B_{\text{end}}$ the forced-answer point of that training trace. The *hold* schedule sets $s = \min(k/b, 1)$; *gate* sets $s = \mathbf{1}[k \geq b]$; *tent* rises as hold does and then falls linearly to zero over the next b tokens. Studies 2 and 3 use hold throughout; Study 1's menus include gate and tent candidates (Table 3).

Losses. With $d(t) = h(t) - \big(h_{\text{base}}(t) + s(t)\,a\,\hat{u}\big)$, full-vector reconstruction averages $\|d(t)\|^2/(a^2 + 10^{-8})$ over think tokens and on-axis reconstruction replaces the numerator with $(d(t)^\top \hat{u})^2$ (Eq. 1). Both therefore use the same target vector and the same schedule. Full-vector reconstruction constrains every component of the activation; on-axis reconstruction constrains only its projection on $\hat{u}$ and adds no penalty on the off-axis components. Because a appears in the target and in the normalizer, changing α changes the target displacement and the loss normalization together, so an α menu is a local recipe search, not a dose series.

Study 1's scalar family also contains targets on the cosine projection $p(t) = h(t)^\top \hat{u}/\|h(t)\|$ with an absolute target m. The *hinge* loss is the mean of $\max(0, m - p)$ over tokens after the cut plus the mean of $\max(0, m + p)$ over tokens before it. The *ramp* loss (candidate S5) averages the smooth-L_1 error between $p(t)$ and $m\,[2\min(k/b, 1) - 1]$. A cosine target m and a reconstruction α are on different scales, so the rows of Table 3 are budget positions, not matched displacements.

Every auxiliary loss has weight 1 and is added to a common completion-only language-model loss on the teacher traces, whose weight after B_{end} is also 1. No backtracking term and no separate off-axis penalty is added in any study.

Menus. Study 1 gives each family seven freshly trained candidates that vary layer, target magnitude, adapter capacity and schedule (Table 3). Its finalists are R2 (full-vector reconstruction, $\alpha = 25$), which was development-feasible, and S7 (on-axis reconstruction, $\alpha = 25$), the scalar family's fallback under the selection rule because no scalar candidate was development-feasible (Section 6.1). Both use layer 18, the hold schedule and attention-only adapters, so S7 shares R2's target exactly and differs only in the loss. S7 is the scalar family's closest member to reconstruction, and a result for S7 does not carry over to the hinge or ramp candidates. Study 2 fits both objectives at $\alpha \in \{5, 10, 15, 20, 25, 35, 50\}$, seven candidates per family named R1 to R7 and S1 to S7 in ascending α; only the loss differs between Rk and Sk. Study 3 uses $\alpha \in \{5, 6, \ldots, 25, 35, 50\}$, 23 candidates per family and 46 in total, named by family and α (R_a05 to R_a50 and S_a05 to S_a50). Its menu contains Study 2's fourteen fitted adapters, mapped to their α values, plus 32 new fits under the same configuration and seed. All 46 candidates and the base model were evaluated afresh on Study 3's 200-problem development panel, which consists of Study 2's 100 development problems followed by 100 new ones; no stored Study 2 generation was reused, although conditions on Study 2's development block reuse its seed, configuration and order and may reproduce previously seen samples.

Adapters and optimization. Table 4 lists the pinned training settings. In every study the writer is layers 0 to L and the reader, layers above L, is frozen; only the final checkpoint of a fit is eligible, and no checkpoint, seed or interpolation search is made. Study 1 records the optimizer as AdamW (the `adamw_torch` default of Transformers 4.48.3) with cosine decay and warmup fraction 0.1. Study 3's fits and the refits ran the trainer through a wrapper that calls `transformers.set_seed`

Table 3: Study 1's frozen seven-candidate menus. A: attention adapters; A+M: attention and MLP adapters. Hinge and ramp m is an absolute cosine target; reconstruction $\alpha/100$ scales the mean activation norm. Each row is a budget position, not an equality of target displacement.

ID	Reconstruction family (R)	Scalar family (S)
1	L18, $\alpha = 10$, hold, A	L18, hinge $m = .15$, A
2	L18, $\alpha = 25$, hold, A	L18, hinge $m = .30$, A
3	L18, $\alpha = 50$, hold, A	L18, hinge $m = .50$, A
4	L24, $\alpha = 25$, hold, A	L24, hinge $m = .50$, A
5	L18, $\alpha = 25$, gate, A	L18, ramp $m = .30$, A
6	L18, $\alpha = 25$, hold, A+M	L18, hinge $m = .50$, A+M
7	L18, $\alpha = 25$, tent, A	L18, on-axis $\alpha = 25$, hold, A

with the training seed at process start and again immediately before `peft.get_peft_model`, so adapter initialization follows the seed; the trainer alone does not seed initialization, and the released code omits this wrapper with the rest of the job-management layer. Study 1's contract has 12 nominal epochs, but its preserved training metadata records 60 optimizer steps, epoch 10.0 and 420 loss calls for each of its 14 candidates, because the formatter accepts 42 of the teacher file's 48 rows, covering 23 of its 24 problem IDs; the other six rows fail the closed-think format check. None of the accepted rows reaches the 10,240-token truncation limit. These counts describe fitting exposure; the nominal 24-problem label alone does not. Studies 2 and 3 use the same archived teacher file and pin the settings in Table 4; we do not report their step counts here.

Table 4: Pinned training settings. Values are protocol parameters, identical across the two families within a study; only the loss (and, in Study 1, the listed menu knobs) differs between candidates.

Setting	Study 1	Studies 2 and 3
Training seed	20260913	20260918
Layer L	18 or 24 (menu)	18
Schedule	hold, gate or tent (menu)	hold
LoRA rank, scaling, dropout	16, 32, 0.05	16, 32, 0.05
Adapted modules	q/k/v/o (+ gate/up/down for A+M)	q/k/v/o
Learning rate	10^{-4}	10^{-4}
Batch size, gradient accumulation	1, 8	1, 8
Nominal epochs	12	12
Maximum sequence length	10,240	10,240
Auxiliary loss weight	1	1
Candidates per family	7	7 (Study 2), 23 (Study 3)
Target magnitude	α or m (menu)	α (menu)

Scope of the comparison. Within a study, both families receive the same number of candidates, the same fitting inputs, seed and development evaluation. This equalizes the incremental search in each study only. Earlier recipe searches were not balanced across families, and equal menus do not undo that. Each study fits one training seed, so its intervals are conditional on the selected checkpoints and carry no training-seed variation.

B Evaluation protocol and decision rules

This appendix states, for each study, the evaluation contract, the estimators, the interval rules, the development selection rule and the decision table that maps intervals to the reported cells. Every rule was fixed before the corresponding held-out outcomes existed; Appendix H records the operational amendments made during collection. Throughout, P means pass, F a resolved failure and U unresolved; U is not a pass, and an unresolved difference is not equivalence.

Table 1 summarizes the three studies.

B.1 DECODING, ROLLOUTS AND GRADERS

Decoding contract. All generation uses vLLM with temperature 0.7, top-p 0.95, at most 16,384 new tokens, four samples per call, tensor parallelism 4, and the model's own end-of-sequence stopping; no stop string or stop-token list is supplied. The single user turn, rendered with the model's chat template, is `Solve this math problem step by step. You MUST put your final answer in \boxed{}. Problem:` followed by the problem text; it is identical across arms, phases and studies. Prompts, problem order and request blocks of 100 problems are identical across arms within a study, and the same decoder, graders and configuration serve development and held-out generation. Study 1 additionally records vLLM 0.6.6, bf16, top-k disabled, maximum model length 17,408, eager execution, and prefix caching and chunked prefill disabled. Engine seeds identify draws: Study 1 used 101 for development and 173 and 271 for its two held-out draws; Study 2 used 101 and 173, 271; Study 3 used 101, 307, 401 and 503 for its four development draws and 173, 271, 619 and 827 for its four held-out draws. Sampling seeds are unset within a call.

Rollout inventories. Each logical job evaluates one arm on one block of 100 problems in one draw, four samples per problem. Study 1 evaluates base and its two finalists on 2,000 held-out problems in two draws, a fixed expected inventory of 120 jobs and 48,000 rows. Study 2 evaluates base and 14 candidates on 100 development groups in one draw, and base and its two held-out arms (the selected pair, or the diagnostic finalists when no pair is eligible) on its certified held-out set of 2,130 problems in two draws. Study 3 evaluates base and all 46 candidates on 200 development groups in four draws (a fixed 376 jobs and 150,400 rows, all generated afresh), then base and the union of the selected arms, at most seven conditions, on 2,000 held-out problems in four draws: at most 224,000 rows in 560 jobs. Capped, unclosed, wrong, malformed and unparseable completions remain in every denominator. Missing or duplicate records, nonfinite fields or a zero base denominator refuse scoring; they are never imputed or dropped.

Graders. Both graders are mandatory in every study, and neither falls back on the other. The configuration below is the one recorded for Studies 2 and 3; Study 1 graded with the same evaluation module under Math-Verify 0.8.0 and SymPy 1.13.1, and its byte-level grader record is not reproduced here.

- *Math-Verify* (version 0.8.0, with SymPy 1.13.1 and antlr4-python3-runtime 4.13.2) compares the gold answer, wrapped in `$...$` unless it already starts with `$` or `\boxed`, with the *full* completion, reasoning included. No box is required. The library's default extraction settings apply (any-match extraction, first-match fallback, extraction without an anchor allowed); it extracts one answer candidate, the last parseable match of the highest-priority pattern class, so an earlier matching expression is not credited when a later candidate parses. Parsing and verification each have a five-second timeout, and an exception or timeout counts as incorrect.
- *Legacy* takes the last `\boxed{...}` in the completion (one level of nested braces allowed), strips `$`, `\left`, `\right` and spacing, and accepts an exact string match with the gold answer or equality after conversion to a number. A completion without a box is incorrect.
- *Closure* is the literal presence of `</think>`. *Think tokens* are the tokenized length of the text before the first `</think>`, or of the whole completion when it does not close.

The two graders differ by construction: Math-Verify can credit an answer written without a box, including one read from inside an unclosed, capped trace, and legacy cannot. Their disagreements are reported as findings (Section 6.4 and Appendix G); no disagreement was used to rescore, select or correct any arm.

B.2 ESTIMATORS

For arm a and problem i, let m_{ai} be the mean over draws of the within-draw median of the four think lengths, and let p_{ai} and h_{ai} be the correctness (under one grader) and closure flags averaged over all draws and samples. With B the base arm,

$$C_a = 100\Big(1 - \frac{\sum_i m_{ai}}{\sum_i m_{\mathrm{B}i}}\Big), \quad D = C_{\mathrm{R}} - C_{\mathrm{S}}, \quad A_{ab} = 100\,\overline{(p_{ai} - p_{bi})}, \quad H_{ab} = 100\,\overline{(h_{ai} - h_{bi})},$$

where the bar is an equal-weight mean over problems. Accuracy contrasts are computed separately under each grader and written, for example, as the Math-Verify R−S or legacy R−B contrast. The medians are within draws, over four samples, never over a pooled set of samples; C_a is a ratio of sums, never a mean of per-problem ratios. In Studies 2 and 3, D is a matching diagnostic with a ± 1 pp tolerance; in Study 1 it is the primary compression contrast, with a minimum interesting magnitude of 5 pp. The two roles are distinct and never combined.

B.3 Intervals

Bootstrap. All three studies use a paired one-stage percentile bootstrap over problems: a replicate resamples whole problems with replacement, and each resampled problem carries every arm, draw and sample, so all contrasts in a replicate share one resample. Execution blocks are not strata. Study 1 uses 200,000 common replicate index vectors with seed 17. Studies 2 and 3 use NumPy 2.2.6 PCG64 with seed 543, 200,000 replicates in batches of 128 and linear quantiles; in Study 3 one set of resamples serves all three targets. In Studies 2 and 3, inferential endpoints are rounded to ten decimals before classification. In every study a value exactly on a boundary is unresolved. A contrast that is constant across problems, or an interval of zero width, is unresolved whatever its endpoints; an empty replicate array or a nonfinite replicate makes the analysis invalid.

Interval families. Table 5 gives the quantiles behind each named bound. Study 3's adjusted levels apply a Bonferroni divisor of 3, one per planned target, to each endpoint family separately, and the three targets always count, whether or not a target was tested. The adjusted display is not a simultaneous 95% guarantee across endpoint families, graders or arms, and the choice between R and S, or among endpoint families, is not covered by any omnibus guarantee. If one pair had been selected at two targets it would be one evaluated pair, reported identically at each, with the divisor still 3; this did not occur. Unadjusted intervals are secondary descriptive columns for Study 3.

Table 5: Quantiles of the bootstrap distribution behind each bound. "One-sided" bounds are used singly, a lower bound for a pass and an upper bound for a failure.

Study	Endpoints	Quantiles
1	C_{R}, C_{S}, D (two-sided 95%)	2.5th and 97.5th percentiles
1	six safeguard contrasts (one-sided 95%)	5th and 95th percentiles
2	accuracy direction and practical magnitude (two-sided 95%)	2.5th and 97.5th percentiles
2	match on D, within-margin accuracy, safeguards, positive cut (one-sided 95%)	5th and 95th percentiles
3	accuracy direction and practical magnitude (two-sided $98\frac{1}{3}\%$)	$q(0.025/3)$ and $q(1-0.025/3)$
3	match on D, within-margin accuracy, safeguards, positive cut (one-sided $98\frac{1}{3}\%$)	$q(0.05/3)$ and $q(1-0.05/3)$

Margins. The accuracy margin is $\delta_A = 2.6$ pp and the closure margin $\delta_H = 1$ pp. Both were fixed before any held-out outcome as tolerances; they are not estimates of training noise and do not establish that a loss within them is harmless. In Studies 2 and 3 the same 2.6 pp value also serves as the arm-versus-arm equivalence radius and as the threshold of practical interest for an accuracy difference. The matching tolerance on D is 1 pp.

B.4 Study 1: equal selection

Development selection. Each family's seven candidates (Table 3) and base are evaluated on the same 100 development groups, four samples each. A candidate is feasible if its Math-Verify accuracy minus base is at least -2.6 pp, its closure minus base is at least 0 and its cut is positive; these are point-estimate rules, not confidence statements. Within each family the finalist maximizes the cut over feasible candidates, breaking exact ties by higher accuracy, then higher closure, then lower candidate number, on unrounded values. If no candidate is feasible, the same ordering applies to all seven and the finalist carries a permanent `NO-DEV-FEASIBLE` flag; it can receive a held-out verdict but never becomes development-qualified. Both menus complete before selection, and held-out outcomes cannot change it.

Held-out states. For a proposed winner a there are six safeguard components: accuracy against base and against the other finalist under Math-Verify and under legacy (margin δ_A), and closure against both comparators (margin δ_H). A component is P if its one-sided lower bound is strictly above the negative margin, F if its upper bound is strictly below it, and U otherwise; the aggregate Q_a is F if any component is F, P if all six are P, and U otherwise. The absolute cut C_a is P if the lower end of its two-sided 95% interval is positive, F if the upper end is negative, and U otherwise. D resolves toward R only when its point and lower endpoint are positive, and toward S only when its point and upper endpoint are negative; otherwise the cell is `COMPRESSION-UNRESOLVED`, whatever either arm's comparison with base.

Cells. Structural states take precedence in the order `NOT-STARTED`, `INVALID`, `INCOMPLETE`, selection failure, then complete; coexisting flags are all kept. For a resolved direction toward arm a, the cell is the prefix R or S followed by the entry of Table 6. A qualifier, `RESOLVED-BELOW-INTEREST`, is added when D resolves and its unrounded point magnitude is strictly below the 5 pp minimum of interest (exactly 5 pp is not tagged); it leaves the cell unchanged and rules out any claim of a practically meaningful difference. The frozen scorer does not emit this qualifier, so a separate hash-bound reporting step applies it to the sealed point and interval.

Table 6: Study 1 cells for a resolved compression direction toward arm a, by a's absolute cut state (rows) and aggregate safeguard state Q_a (columns). † adds "excessive quality loss" and ‡ adds "quality unresolved"; "same" repeats `ABSOLUTE-UNRESOLVED`.

Absolute cut	$Q_a = \mathrm{P}$	$Q_a = \mathrm{F}$	$Q_a = \mathrm{U}$
P	`WIN`	`TRADEOFF`	`LENGTH-ONLY`
F	`NO-HALT`	`NO-HALT`†	`NO-HALT`‡
U	`ABSOLUTE-UNRESOLVED`	same†	same‡

Replacement and stopping. One operational replacement is allowed per logical held-out job, only for a documented technical failure; the earliest complete valid attempt is adopted and failed attempts are preserved. Long reasoning, cap hits, low accuracy or failed scientific criteria never justify a replacement. Missing required data at the stop yields `INCOMPLETE`, with no reduced-sample verdict and no imputation.

B.5 Study 2: matched natural shortening

Development selection. Base and the 14 candidates are evaluated on 100 development groups, four samples each, in one draw. A candidate is feasible if both its Math-Verify and its legacy accuracy minus base are at least -2.6 pp, its closure minus base is at least 0 and its cut is positive. A pair (r, s) is eligible if both are feasible, $\max(|C_r - 25|, |C_s - 25|) \le 2.5$ and $|C_r - C_s| \le 1$. The selected pair is the lexicographic minimum over eligible pairs of $\big(\max(|C_r-25|, |C_s-25|),\ |C_r-C_s|,\ -(\mathrm{MV}_r+\mathrm{MV}_s),\ -(\mathrm{legacy}_r+\mathrm{legacy}_s),\ -(H_r+H_s),$ hash of the sorted pair of candidate IDs$\big)$, where MV and legacy are the development Math-Verify and legacy accuracies and H is the development closure rate. Cuts and pp comparisons are rounded to ten decimals and the closure feasibility difference to twelve; tie-breaking uses unrounded rates. Point matching at development is a heuristic, not evidence of equivalence.

No development match. If no pair is eligible the study is permanently `NO-DEV-MATCH:DIAGNOSTIC-ONLY`. Each family then has a diagnostic finalist, chosen from its feasible candidates if any and otherwise from all seven, by closest cut to 25%, then higher Math-Verify accuracy, legacy accuracy and closure, then the hash of the candidate ID; `NO-DEV-FEASIBLE` flags are kept. The diagnostic finalists are evaluated on the whole held-out set, and even a held-out match cannot rescue a failed development match.

Held-out states. The predicates are those of Table 7 with Study 2's unadjusted levels (Table 5), and the cells are those of Appendix B.7.

B.6 STUDY 3: FINER MATCHED FOLLOW-UP

Frame. Study 3 was designed after Study 2's development outcomes were seen and was frozen after Study 2's held-out results, including the diagnostic finalists' held-out cuts, were known, so it is an outcome-informed follow-up with prospectively frozen new evaluation. Regeneration reuses no stored Study 2 rows, but fifteen conditions on the original development block reuse the earlier seed, configuration and order and may reproduce previously seen samples; it removes neither prior exposure nor knowledge of Study 2's candidates. Study 2 and its cell are unchanged.

Development selection. All 47 conditions (base and 46 candidates) are evaluated on the same 200 development groups (Study 2's 100 development groups, then 100 groups from the population Study 3 adopted) in four draws before any target is selected. Feasibility is Study 2's expression on all 200 groups. The three targets are cuts of 12.5, 17.5 and 22.5% of base think tokens, with inclusive windows $[10, 15]$, $[15, 20]$ and $[20, 25]\%$ after rounding. For each target, all $23 \times 23 = 529$ pairs are enumerated; a pair is eligible if both candidates are feasible, both cuts lie in the window and the rounded $|D|$ is at most 1 pp. A candidate or pair on a shared window boundary is eligible for both adjacent targets and is never reassigned. Each target independently takes the lexicographic minimum of Study 2's six-part key with 25 replaced by that target, with no global optimization, unique-pair constraint or best-target choice; the pair hash is SHA-256 of the canonical JSON of the sorted candidate IDs. A target with no eligible pair is permanently `NO-DEV-MATCH:UNTESTED` and, unlike Study 2, receives no diagnostic held-out evaluation. No development enlargement, extra draw, α, target, tolerance or held-out top-up is allowed after development results.

Held-out labels. The values 12.5, 17.5 and 22.5 label development selection only. Held-out cuts are reported as observed, with their signed deviation from the label, and held-out window adherence is not a test gate. Two tested targets are called distinct held-out operating points only descriptively, and only if no unadjusted 95% interval for C_{R} or C_{S} at one target overlaps either interval at the other (all four comparisons; closed intervals, so touching counts as overlap). Overlap does not establish equality, and no paired contrast between targets was specified or is computed.

Held-out states. The predicates are those of Table 7 with Study 3's adjusted levels, and the cells are those of Appendix B.7. In Study 3's frozen analysis, the `SHORTENING-NOT-ESTABLISHED` prefix is applied at targets where matching is established and either absolute cut lacks P.

Replacement. Every fit and every development or held-out job (one arm, one block, one draw) allows at most two attempts with identical configuration, the second only after a technical failure. The first valid attempt is retained, all attempts and adverse completions are kept, and a poor result is never retried.

B.7 HELD-OUT STATES AND CELLS FOR STUDIES 2 AND 3

Table 7: Held-out predicates for Studies 2 and 3. "Lower" and "upper" are the one-sided bounds and "2-sided" the two-sided interval of Table 5 (unadjusted in Study 2, adjusted in Study 3). A degenerate contrast is U in every row. Direction takes precedence over E.

Endpoint	P or R	F or S	Otherwise
Match on D	lower > -1 and upper < 1: P	lower > 1 or upper < -1: F	U
Accuracy R−S, each grader	2-sided lower > 0: R	2-sided upper < 0: S	E if lower > -2.6 and upper < 2.6; else U
Safeguard component	lower $> -\delta$: P	upper $< -\delta$: F	U
Absolute cut C_a	lower > 0: P	upper < 0: F	U

Safeguards. Each arm has six components: Math-Verify accuracy, legacy accuracy ($\delta = \delta_A$) and closure ($\delta = \delta_H$), each against base and against the other arm. Q_a is F if any component is F, P if all six are P, and U otherwise. Both Q values and all twelve components are reported, not only those of a winning arm. A passing safeguard is not accuracy preservation.

Cells. Structural states come first: NOT-STARTED, INVALID (proven leakage, asymmetric opportunity, corrupt identity or producer), INCOMPLETE (missing required inventory or receipt), then the development outcome (NO-DEV-MATCH:DIAGNOSTIC-ONLY in Study 2, NO-DEV-MATCH:UNTESTED per target in Study 3). Otherwise, a match state of F gives MATCH-FAILED and U gives MATCH-UNRESOLVED, whatever the accuracy results. Only a match state of P enters Table 8. A resolved direction toward arm a under both graders becomes a-ACCURACY-ADVANTAGE if Q_a is P, a-ACCURACY-TRADEOFF if F, and a-ACCURACY-ONLY-QUALITY-UNRESOLVED if U. If either absolute cut lacks P, the prefix SHORTENING-NOT-ESTABLISHED is added and no preserved-shortening wording is allowed.

Table 8: Cells at an established match, by Math-Verify state (rows) and legacy state (columns). R and S: resolved direction; E: within the 2.6 pp margin; U: unresolved. ACC. abbreviates ACCURACY.

MV \ legacy	R	S	E	U
R	R direction	GRADER-SPLIT	GRADER-SPLIT	GRADER-SPLIT
S	GRADER-SPLIT	S direction	GRADER-SPLIT	GRADER-SPLIT
E	GRADER-SPLIT	GRADER-SPLIT	WITHIN-ACC.-MARGIN	GRADER-SPLIT
U	GRADER-SPLIT	GRADER-SPLIT	GRADER-SPLIT	ACC.-UNRESOLVED

Reading the cells. WITHIN-ACCURACY-MARGIN means that, under both graders, no direction is resolved and both one-sided bounds on the R−S accuracy contrast lie strictly inside ± 2.6 pp. It argues against a difference larger than the margin, but it is not equality and does not certify retention of base accuracy. ACCURACY-UNRESOLVED supports neither equality nor an advantage. GRADER-SPLIT means the conclusion depends on the grader. MATCH-FAILED and MATCH-UNRESOLVED cannot support an accuracy comparison at matched shortening, even when an accuracy difference is resolved. NO-DEV-MATCH reports that the menu contained no eligible pair, not that no matching pair exists elsewhere.

Qualifiers and flags. These are reported beside the cell and never change it. For each grader with a resolved direction, the magnitude is PRACTICALLY-ESTABLISHED if the directional lower end of its two-sided interval exceeds 2.6 pp, BELOW-INTEREST if the directional upper end is below 2.6 pp, and INTEREST-UNRESOLVED otherwise (for a direction toward S, the interval is sign-flipped first). The GRADER-SPLIT flag is set when the two graders' direction states differ or when any safeguard component differs between Math-Verify and legacy. The DRAW-SIGN-SPLIT flag is set when the per-draw points of D, the Math-Verify or legacy R−S contrast, or the closure R−S contrast take opposite signs across draws. With several draws (two in Study 2, four in Study 3) and a small true difference, it is expected to fire often; it never overrides a verdict. Per-draw states are descriptive summaries, not tests: the draws share the same problems and bootstrap seed (Appendix F).

C STUDY 1 FULL RESULTS

Finalists and inventory. The two seven-candidate menus are in Appendix A. The full-vector finalist R2 was development-feasible. No scalar candidate was, so the scalar family sent its fallback S7, whose NO-DEV-FEASIBLE flag is permanent: S7 receives a held-out cell but cannot become development-qualified. S7 is the scalar family's closest member to reconstruction, sharing R2's target and schedule, so a result for S7 says nothing in favor of the hinge or ramp targets. All 120 fixed held-out jobs and 48,000 rows over 2,000 problem groups were received. The jobs took 121 attempts: the extra attempt replaced a documented technical failure under the fixed replacement rule, and the failed attempt is retained. The admission stop was extended by one day after a validation failure and before any held-out job ran.

Every component, combined and by draw. Table 9 reports D, both absolute cuts and all six quality components of each finalist for the combined analysis and for each draw. The recorded flags are: DRAW-SIGN-SPLIT:closure_RS, DRAW-SIGN-SPLIT:mv_RB, GRADER-SPLIT, S:NO-DEV-FEASIBLE, draw=1:GRADER-SPLIT, draw=2:GRADER-SPLIT. GRADER-SPLIT marks components on which the two graders reach different states: S7 against base and S7 against R2. DRAW-SIGN-SPLIT marks the two components whose draw points have opposite signs, R2's

closure against S7 and R2's Math-Verify accuracy against base. The draws sample the same problems, so the per-draw columns are descriptive and add no independent test.

Table 9: Study 1 held-out results, combined and by draw. Estimates and bounds are percentage points, except cuts (%). Brackets give two-sided 95% intervals for D and the cuts, and one-sided 95% bounds (5th and 95th percentiles) for quality components, judged against -2.6 pp (accuracy) and -1.0 pp (closure). MV is Math-Verify. States are copied from the final seal: S means D resolves toward S7; P, F and U mean pass, fail and unresolved. Every analysis carries `RESOLVED-BELOW-INTEREST`. Draw columns are descriptive.

	Combined			Draw one			Draw two		
Component	Est.	Bounds	St.	Est.	Bounds	St.	Est.	Bounds	St.
D, R2 minus S7	−3.29	[−4.20, −2.37]	S	−3.61	[−4.82, −2.39]	S	−2.97	[−4.16, −1.77]	S
R2 absolute cut (%)	26.58	[25.57, 27.57]	P	26.16	[24.85, 27.43]	P	27.00	[25.71, 28.28]	P
accuracy vs base, MV	−0.19	[−0.73, +0.34]	P	−0.59	[−1.31, +0.14]	P	+0.20	[−0.49, +0.89]	P
accuracy vs S7, MV	+1.93	[+1.43, +2.43]	P	+1.66	[+0.98, +2.35]	P	+2.20	[+1.55, +2.86]	P
accuracy vs base, legacy	−0.74	[−1.26, −0.22]	P	−1.18	[−1.90, −0.45]	P	−0.30	[−0.99, +0.38]	P
accuracy vs S7, legacy	+6.19	[+5.59, +6.79]	P	+5.88	[+5.08, +6.68]	P	+6.50	[+5.74, +7.28]	P
closure vs base	+6.15	[+5.59, +6.72]	P	+6.00	[+5.34, +6.68]	P	+6.30	[+5.66, +6.95]	P
closure vs S7	−0.28	[−0.57, 0.00]	P	−0.58	[−0.94, −0.21]	P	+0.01	[−0.35, +0.36]	P
aggregate Q_{R2}	n/a	n/a	P	n/a	n/a	P	n/a	n/a	P
S7 absolute cut (%)	29.87	[28.74, 30.97]	P	29.77	[28.38, 31.12]	P	29.97	[28.66, 31.25]	P
accuracy vs base, MV	−2.13	[−2.74, −1.51]	U	−2.25	[−3.01, −1.48]	U	−2.00	[−2.76, −1.24]	U
accuracy vs R2, MV	−1.93	[−2.43, −1.43]	P	−1.66	[−2.35, −0.98]	P	−2.20	[−2.86, −1.55]	U
accuracy vs base, legacy	−6.93	[−7.64, −6.22]	F	−7.05	[−7.93, −6.19]	F	−6.80	[−7.68, −5.95]	F
accuracy vs R2, legacy	−6.19	[−6.79, −5.59]	F	−5.88	[−6.68, −5.08]	F	−6.50	[−7.28, −5.74]	F
closure vs base	+6.43	[+5.83, +7.05]	P	+6.58	[+5.88, +7.29]	P	+6.29	[+5.61, +6.98]	P
closure vs R2	+0.28	[0.00, +0.57]	P	+0.58	[+0.21, +0.94]	P	−0.01	[−0.36, +0.35]	P
aggregate Q_{S7}	n/a	n/a	F	n/a	n/a	F	n/a	n/a	F

Arm levels. Table 10 gives row-level means over the 16,000 held-out rows of each arm. They are descriptive summaries, not the frozen per-problem estimators, and they are not comparable with the levels of Studies 2 and 3, which use other populations. Appendix G compares the two graders' behavior.

Table 10: Study 1 held-out arm levels (row means over both draws; descriptive).

Arm	Think tokens	Math-Verify	Legacy	Closure	Capped
Base	5,628	0.819	0.713	0.919	0.087
R2 (full-vector)	4,208	0.817	0.706	0.981	0.021
S7 (on-axis)	4,027	0.798	0.644	0.984	0.017

Scope and disclosures. Early-era project exposure was not certified. The population enforced modern-era exposure exclusions and disjointness of fitting, development and held-out inputs; this is bounded separation from project use, not a certificate of clean pretraining. Equal selection equalizes only the incremental search: earlier recipe searches were not balanced across families. Inference concerns two frozen checkpoints, one training seed per candidate and two evaluation draws; it says nothing about the reproducibility of the training procedure, and the study makes no necessity, causal, optimality, efficiency-frontier or equivalence claim. Sample-size planning beyond the observed panels used conditional simulations, not measured power. The frozen scorer does not emit the `RESOLVED-BELOW-INTEREST` tag, so a separate reporting step, bound to the sealed analysis by hash, applied it to the exact sealed point and interval.

Cross-check and provenance. A text-derived reimplementation of the analysis was frozen before the production analysis source was inspected and then compared with it on prespecified synthetic cases. All numeric fields agreed; the cell strings agreed only after omitted quality qualifiers were repaired against the reference, so their final agreement is post-repair consistency, not independent confirmation. On the real data, the independent analysis agrees on numerical estimates, intervals, cells and inventory but disagrees on the `RESOLVED-BELOW-INTEREST` tag for the combined analysis and both draws. Continuation was authorized with that disagreement retained and the separate

qualifier applied, so this is not an unqualified agreement pass. Among the base-model jobs, 3 lack the external compute provider's return record and are admitted using campaign-generated HMAC bindings and retained artifacts on the same storage volume. This establishes consistency within that trust boundary, not independent provider authentication or protection against the shared storage administrator.

D STUDY 2 FULL RESULTS

Development and pair classification. Table 11 gives each candidate's development cut, its grader and closure counts, and their differences against base, on 100 problem groups with four samples each. Of 7 candidates per family, 5 full-vector and 2 on-axis candidates were feasible. All infeasible candidates fail legacy retention; R6, R7, S6 and S7 also fail Math-Verify retention, and S7 also fails closure. These are development point predicates, not confidence statements. Table 12 classifies the 49 pairs of the fixed seven-by-seven menu. The per-arm window (both cuts in $[22.5, 27.5]\%$) fails for 47 pairs and the between-arm tolerance ($|D| \leq 1$ pp) for 46, so the shortening restrictions alone leave no pair; feasibility separately excludes 39. Every pair of two feasible candidates fails a shortening restriction. Only R5 and S2 are feasible with both cuts in the window, and they differ by $+2.50$ pp. No near miss changes the rule, and the empty set makes `NO-DEV-MATCH:DIAGNOSTIC-ONLY` permanent.

Table 11: Study 2 development results for base and the 14 candidates (100 groups, four samples each). Cuts are percent of base think tokens; counts are out of 400 completions; differences against base are percentage points. Feasible means both accuracy differences ≥ -2.6 pp, a closure difference ≥ 0 and a positive cut.

			Counts (of 400)			Difference vs base (pp)			
Candidate	α	Cut (%)	MV	legacy	closure	MV	legacy	closure	Feasible
Base	n/a	n/a	325	282	364	n/a	n/a	n/a	n/a
R1	5	5.80	326	282	371	+0.25	0.00	+1.75	yes
R2	10	11.03	323	281	376	−0.50	−0.25	+3.00	yes
R3	15	18.46	324	282	384	−0.25	0.00	+5.00	yes
R4	20	21.28	316	273	389	−2.25	−2.25	+6.25	yes
R5	25	25.41	324	281	387	−0.25	−0.25	+5.75	yes
R6	35	28.45	314	266	392	−2.75	−4.00	+7.00	no
R7	50	30.97	305	258	387	−5.00	−6.00	+5.75	no
S1	5	9.81	323	276	384	−0.50	−1.50	+5.00	yes
S2	10	22.90	325	273	389	0.00	−2.25	+6.25	yes
S3	15	20.70	315	259	391	−2.50	−5.75	+6.75	no
S4	20	27.73	328	268	392	+0.75	−3.50	+7.00	no
S5	25	24.23	319	266	388	−1.50	−4.00	+6.00	no
S6	35	34.69	305	252	384	−5.00	−7.50	+5.00	no
S7	50	27.59	292	238	362	−8.25	−11.00	−0.50	no

Table 12: Classification of the 49 development pairs of Study 2. "Shortening" combines the per-arm window and the between-arm tolerance. The two exclusions overlap; the table does not say that both are needed to empty the set.

Class	Pairs	Meaning
Eligible	0	feasible, and both shortening restrictions pass
Feasibility fails only	0	shortening passes, feasibility fails
Shortening fails only	10	both feasible, a shortening restriction fails
Both fail	39	feasibility and a shortening restriction fail

Held-out results. Table 13 reports every held-out component for the diagnostic finalists on 2,130 problem groups (51,120 fixed expected rows; two draws of four samples). The matching component is

F: the one-sided lower bound of D lies above $+1$ pp. Both finalists pass all six safeguards (aggregate states P and P); this is not accuracy preservation, and the accuracy points against base are not bounds on possible loss. On R−S accuracy, the Math-Verify contrast resolves toward S2 (state S) and the legacy contrast toward R5 (state R), each below the 2.6 pp practical threshold, and the analysis records `GRADER-SPLIT`. Under `NO-DEV-MATCH` these contrasts are diagnostic and do not rank the objectives. Table 14 gives the arm levels and each draw's points (engine seeds 173 and 271). The draws sample the same problems, so the draw rows are descriptive, and the levels are not comparable with those of Studies 1 and 3.

Table 13: Study 2 held-out results for the diagnostic finalists R5 (full-vector, $\alpha = 25$) and S2 (on-axis, $\alpha = 10$). Cuts are percent of base think tokens; other rows are percentage points. Two-sided 95% intervals are the direction and descriptive bounds; one-sided 95% bounds (5th and 95th percentiles) test matching against ± 1 pp, accuracy safeguards against -2.6 pp and closure safeguards against -1.0 pp. States: absolute-cut state for the cuts, matching state for D, safeguard state for the other rows.

Endpoint	Point	Two-sided 95%	One-sided 95% bounds	State
C_R (R5 cut, %)	26.78	[25.78, 27.77]	[25.94, 27.62]	P
C_S (S2 cut, %)	22.52	[21.48, 23.54]	[21.65, 23.38]	P
$D = C_R - C_S$	+4.26	[+3.46, +5.07]	[+3.59, +4.94]	F
MV, R5 − base	−0.60	[−1.20, −0.01]	[−1.10, −0.10]	P
legacy, R5 − base	−0.87	[−1.51, −0.23]	[−1.40, −0.33]	P
closure, R5 − base	+6.38	[+5.73, +7.06]	[+5.83, +6.95]	P
MV, S2 − base	+0.38	[−0.22, +0.99]	[−0.13, +0.89]	P
legacy, S2 − base	−1.56	[−2.27, −0.85]	[−2.15, −0.96]	P
closure, S2 − base	+6.47	[+5.82, +7.14]	[+5.92, +7.04]	P
MV, R5 − S2	−0.98	[−1.49, −0.47]	[−1.41, −0.55]	P
legacy, R5 − S2	+0.69	[+0.15, +1.23]	[+0.23, +1.14]	P
closure, R5 − S2	−0.09	[−0.35, +0.17]	[−0.31, +0.13]	P
MV, S2 − R5	+0.98	[+0.47, +1.49]	[+0.55, +1.41]	P
legacy, S2 − R5	−0.69	[−1.23, −0.15]	[−1.14, −0.23]	P
closure, S2 − R5	+0.09	[−0.17, +0.35]	[−0.13, +0.31]	P

Table 14: Study 2 held-out arm levels (combined over both draws; points only) and each draw's points for the R5 minus S2 contrasts (percentage points; descriptive).

Arm	Think tokens	Math-Verify	Legacy	Closure
Base	5,526	0.807	0.691	0.917
R5 (full-vector)	4,046	0.801	0.682	0.981
S2 (on-axis)	4,282	0.811	0.675	0.982
Draw	D	MV, R5−S2	Legacy, R5−S2	Closure, R5−S2
One	+4.30	−0.76	+0.86	−0.11
Two	+4.22	−1.20	+0.52	−0.07

Disclosures. The frozen scorer bounded re-encoded think lengths by the 16,384-token generation cap. Of the 51,120 held-out rows, 3 re-encode to one token more than the cap although exactly 16,384 tokens were generated, because decoding and re-encoding can change token merges, and the frozen scorer's validation refused them. After that validation failure and before any aggregate outcome was inspected, the scorer was amended as a versioned successor: it accepts any non-negative integer re-encoded length, uniformly across arms and phases, checks the generated-token count against the cap separately, and clamps, excludes and regenerates nothing. The original scorer, its failed run and the affected rows are preserved, and the development results and finalist selection are unchanged. No standalone runtime acceptance record for the grader predates development: the sealed selection designates in-image positive and negative controls and pinned grader bytes, and a replay after the seal reproduced all 18,000 expected grader flags on the development rows. That replay is neither an independent mathematical adjudication nor a retrospective acceptance.

E STUDY 3 FULL RESULTS

This appendix reports Study 3's development selection record and every held-out endpoint of the frozen analysis, for all three targets; per-recipe development formatting counts are in Appendix N. Development statistics are selection inputs, not held-out results. Held-out intervals come from the paired problem bootstrap of Section 3; "adjusted" means Bonferroni over the three planned targets within an endpoint family (one-sided or two-sided $98\frac{1}{3}\%$), and unadjusted two-sided 95% intervals are secondary columns. No interval is a simultaneous guarantee across endpoint families, graders or arms. We write R_α and S_α for full-vector and on-axis reconstruction at recipe parameter α, and MV for Math-Verify; a contrast such as MV, R−S is the full-vector minus on-axis Math-Verify accuracy in pp.

E.1 DEVELOPMENT SELECTION

Each of the 23 recipes per family, and the base model, was sampled on 200 development groups in four draws of four. A recipe is feasible if its Math-Verify and legacy accuracy point differences against the base model are both at least -2.6 pp, its closure difference is not negative, and its cut is positive. For each target, a pair of the 529 is eligible if both recipes are feasible, both development cuts lie in the target's window, and $|D| \leq 1$ pp; the frozen lexicographic key then selects one eligible pair per target, independently across targets. These are point-selection rules on development data, not confidence statements. Table 15 gives the outcome per target, Table 16 every candidate, and Figure 2 the menu with the three windows.

Table 15: Study 3 development selection per target: eligible pairs of 529 and the number of pairs failing each eligibility predicate (a pair can fail several). Window: the recipe's development cut lies in the target window; gap: $|D| \leq 1$ pp.

			Pairs failing				
Target (window)	Selected pair	Eligible	R feasible	S feasible	R window	S window	gap
12.5 ([10, 15])	R_{10} / S_6	3	46	368	414	483	487
17.5 ([15, 20])	R_{14} / S_8	6	46	368	391	460	487
22.5 ([20, 25])	R_{21} / S_{11}	6	46	368	368	368	487

E.2 HELD-OUT ENDPOINTS

Tables 17, 18 and 19 give all fifteen endpoints at each target. The match state uses the adjusted one-sided bounds on D; accuracy direction and practical magnitude use the adjusted two-sided intervals of the R−S contrasts; quality components and absolute cuts use adjusted one-sided bounds. The S−R rows mirror the R−S rows and enter the on-axis arm's safeguards. The frozen states, each listed at targets 12.5, 17.5 and 22.5 in that order, are: match U, U and P; Math-Verify accuracy S, E and E; legacy accuracy E, R and R. Practical-magnitude qualifiers: Math-Verify BELOW-INTEREST, NO-RESOLVED-DIRECTION and NO-RESOLVED-DIRECTION; legacy NO-RESOLVED-DIRECTION, BELOW-INTEREST and INTEREST-UNRESOLVED. Cells: MATCH-UNRESOLVED, MATCH-UNRESOLVED and GRADER-SPLIT. Flags: GRADER-SPLIT, DRAW-SIGN-SPLIT at 12.5, GRADER-SPLIT, DRAW-SIGN-SPLIT at 17.5 and GRADER-SPLIT, DRAW-SIGN-SPLIT at 22.5; the draw-sign flag names the endpoints whose four per-draw points take both signs (12.5: D; 17.5: mv_RS; 22.5: D, mv_RS). Observed held-out cuts deviate from their development labels by $+2.17$ (R) and $+1.84$ (S) at 12.5, $+1.37$ and $+2.20$ at 17.5, and $+1.06$ and $+0.89$ at 22.5 (pp, descriptive; window adherence is not a held-out gate).

Quality components. Each selected arm faces six safeguards: Math-Verify accuracy, legacy accuracy and closure, each against the base model and against the other arm of its pair. A component passes (P) if its adjusted one-sided lower bound exceeds the negative margin, fails (F) if its upper bound lies below it, and is otherwise unresolved (U); the aggregate is F if any component fails, P if all pass, and U otherwise. All 36 components, with their points, bounds and states, are the arm-minus-base and arm-minus-arm rows of Tables 17 to 19, and the aggregates are in Table 2. A quality pass is not accuracy preservation, and point estimates are not bounds on possible loss.

Table 16: Study 3 development menu: development cut (percent of base think tokens), feasibility, and the feasibility checks each infeasible recipe failed (mv: Math-Verify accuracy; legacy: legacy accuracy; closure). 21 of 23 full-vector and 7 of 23 on-axis recipes are feasible. Development statistics only.

	Full-vector reconstruction			On-axis reconstruction		
α	Dev cut (%)	Feasible	Failed	Dev cut (%)	Feasible	Failed
5	6.25	yes	n/a	10.43	yes	n/a
6	9.32	yes	n/a	13.16	yes	n/a
7	10.25	yes	n/a	15.27	yes	n/a
8	10.37	yes	n/a	17.86	yes	n/a
9	11.91	yes	n/a	19.60	yes	n/a
10	12.45	yes	n/a	20.16	yes	n/a
11	14.47	yes	n/a	22.69	yes	n/a
12	15.22	yes	n/a	22.68	no	legacy
13	15.29	yes	n/a	22.03	no	legacy
14	17.47	yes	n/a	23.29	no	legacy
15	18.01	yes	n/a	24.14	no	legacy
16	18.51	yes	n/a	24.08	no	legacy
17	19.99	yes	n/a	25.63	no	legacy
18	20.13	yes	n/a	25.86	no	legacy
19	20.99	yes	n/a	26.46	no	legacy
20	21.93	yes	n/a	25.80	no	legacy
21	22.49	yes	n/a	27.23	no	legacy
22	22.84	yes	n/a	25.67	no	legacy
23	23.48	yes	n/a	26.60	no	legacy
24	24.99	yes	n/a	26.41	no	legacy
25	25.33	yes	n/a	27.82	no	legacy
35	31.25	no	legacy	29.20	no	mv, legacy
50	31.56	no	mv, legacy	25.87	no	mv, legacy, closure

E.3 Held-out operating points

Two tested targets are called distinct held-out operating points only if none of their four between-target cut comparisons (R−R, S−S, R−S, S−R) has overlapping unadjusted 95% intervals; closed intervals that touch count as overlapping. This is a conservative descriptive criterion fixed before the held-out readout, not a paired contrast; no paired contrast between targets was prespecified or computed, and overlapping intervals would be reported as not distinguished, never as equal. 12 of the 12 comparisons are non-overlapping, so every pair of tested targets is distinct under this criterion.

E.4 Arm levels

Table 20 gives descriptive held-out levels per arm: think length is the mean over problems of the mean of the four within-draw medians, and accuracy and closure are proportions averaged over each problem's sixteen outputs and then equally over problems. The inferential contrasts are the endpoints above. These levels belong to Study 3's panel and are not comparable with Study 1 or Study 2 levels.

E.5 Grader disagreement counts

Table 21 counts outputs by grader agreement for every arm, over all 32,000 outputs per arm; every output is retained under both graders. Appendix G defines the categories and gives the descriptive decomposition.

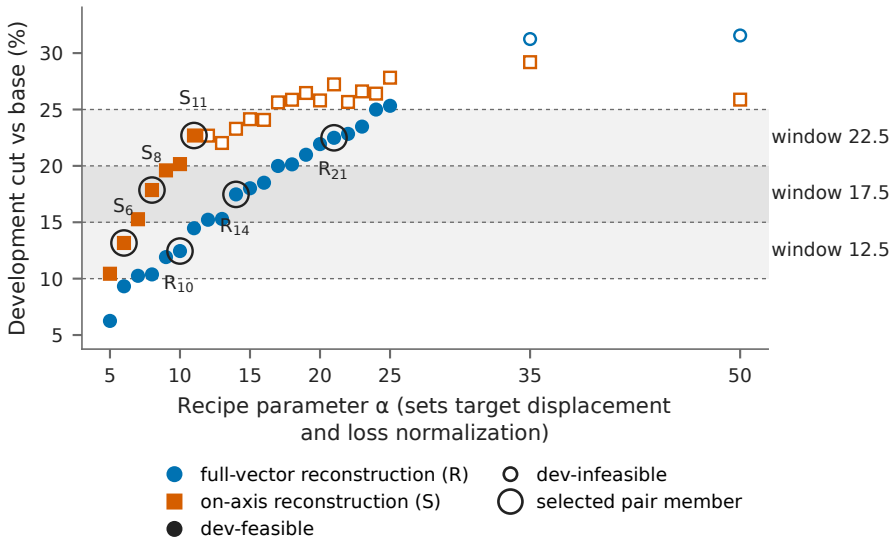

Figure 2: Study 3 development menu (development statistics only). Each marker is one of the 23 recipes per objective, placed by its recipe parameter α and its development cut relative to the base model. Open markers failed the development feasibility rule; every failure involves legacy-graded retention. Shaded bands are the prespecified target windows. A pair was eligible when both members were feasible, both cuts fell in the window and the cuts differed by at most 1 pp; circled markers are the pair the frozen tie-break selected at each target. Because α changes both the target displacement and the loss normalization, the horizontal axis indexes recipes and is not a pure strength.

Table 17: Study 3 held-out endpoints at target 12.5 (R_{10} against S_6), in pp; cuts in percent of base think tokens. Adjusted intervals are primary; the unadjusted 95% interval is secondary. State: positive-cut state for C_R and C_S, match state for D, and quality-component state for each contrast (arm minus base, or arm minus the other arm).

		Adjusted (98.33%)		Unadjusted 95%	
Endpoint	Point	one-sided bounds	two-sided	two-sided	State
C_R (R_{10} cut, %)	14.67	[13.83, 15.50]	[13.73, 15.60]	[13.90, 15.43]	P
C_S (S_6 cut, %)	14.34	[13.52, 15.15]	[13.42, 15.25]	[13.59, 15.09]	P
$D = C_R - C_S$	+0.33	[−0.41, +1.08]	[−0.51, +1.17]	[−0.35, +1.02]	U
MV, R_{10} − base	+0.22	[−0.25, +0.68]	[−0.31, +0.74]	[−0.21, +0.65]	P
legacy, R_{10} − base	+0.60	[+0.14, +1.06]	[+0.08, +1.12]	[+0.18, +1.02]	P
closure, R_{10} − base	+4.60	[+4.07, +5.14]	[+4.01, +5.22]	[+4.11, +5.10]	P
MV, S_6 − base	+0.72	[+0.25, +1.19]	[+0.19, +1.25]	[+0.29, +1.15]	P
legacy, S_6 − base	+0.91	[+0.44, +1.37]	[+0.38, +1.43]	[+0.48, +1.34]	P
closure, S_6 − base	+4.79	[+4.25, +5.35]	[+4.18, +5.43]	[+4.29, +5.31]	P
MV, R_{10} − S_6	−0.50	[−0.93, −0.07]	[−0.98, −0.02]	[−0.89, −0.11]	P
legacy, R_{10} − S_6	−0.31	[−0.73, +0.12]	[−0.79, +0.17]	[−0.70, +0.08]	P
closure, R_{10} − S_6	−0.19	[−0.49, +0.11]	[−0.53, +0.14]	[−0.47, +0.08]	P
MV, S_6 − R_{10}	+0.50	[+0.07, +0.93]	[+0.02, +0.98]	[+0.11, +0.89]	P
legacy, S_6 − R_{10}	+0.31	[−0.12, +0.73]	[−0.17, +0.79]	[−0.08, +0.70]	P
closure, S_6 − R_{10}	+0.19	[−0.11, +0.49]	[−0.14, +0.53]	[−0.08, +0.47]	P

E.6 Per-problem shortening at 22.5 (post hoc, exploratory)

Matching in Study 3 concerns aggregate thinking length, not which problems each arm shortens. This analysis was specified after one held-out draw had been inspected (with a second as reference) and before it was extended to all four draws; it tests nothing and changes no estimator, verdict or reported

Table 18: Study 3 held-out endpoints at target 17.5 (R_{14} against S_8), in pp; cuts in percent of base think tokens. Adjusted intervals are primary; the unadjusted 95% interval is secondary. State: positive-cut state for C_R and C_S, match state for D, and quality-component state for each contrast (arm minus base, or arm minus the other arm).

		Adjusted (98.33%)		Unadjusted 95%	
Endpoint	Point	one-sided bounds	two-sided	two-sided	State
C_R (R_{14} cut, %)	18.87	[18.05, 19.70]	[17.95, 19.80]	[18.11, 19.63]	P
C_S (S_8 cut, %)	19.70	[18.75, 20.64]	[18.64, 20.75]	[18.83, 20.56]	P
$D = C_R - C_S$	−0.83	[−1.61, −0.03]	[−1.71, +0.07]	[−1.55, −0.10]	U
MV, R_{14} − base	+0.27	[−0.22, +0.77]	[−0.28, +0.83]	[−0.18, +0.73]	P
legacy, R_{14} − base	+0.69	[+0.21, +1.18]	[+0.15, +1.25]	[+0.25, +1.15]	P
closure, R_{14} − base	+5.87	[+5.26, +6.50]	[+5.18, +6.58]	[+5.30, +6.45]	P
MV, S_8 − base	+0.46	[−0.06, +0.98]	[−0.12, +1.04]	[−0.02, +0.93]	P
legacy, S_8 − base	−0.90	[−1.49, −0.31]	[−1.57, −0.23]	[−1.45, −0.35]	P
closure, S_8 − base	+6.80	[+6.10, +7.51]	[+6.02, +7.60]	[+6.16, +7.45]	P
MV, R_{14} − S_8	−0.19	[−0.63, +0.25]	[−0.68, +0.30]	[−0.59, +0.22]	P
legacy, R_{14} − S_8	+1.59	[+1.08, +2.11]	[+1.02, +2.17]	[+1.13, +2.07]	P
closure, R_{14} − S_8	−0.92	[−1.22, −0.64]	[−1.25, −0.60]	[−1.19, −0.66]	U
MV, S_8 − R_{14}	+0.19	[−0.25, +0.63]	[−0.30, +0.68]	[−0.22, +0.59]	P
legacy, S_8 − R_{14}	−1.59	[−2.11, −1.08]	[−2.17, −1.02]	[−2.07, −1.13]	P
closure, S_8 − R_{14}	+0.92	[+0.64, +1.22]	[+0.60, +1.25]	[+0.66, +1.19]	P

Table 19: Study 3 held-out endpoints at target 22.5 (R_{21} against S_{11}), in pp; cuts in percent of base think tokens. Adjusted intervals are primary; the unadjusted 95% interval is secondary. State: positive-cut state for C_R and C_S, match state for D, and quality-component state for each contrast (arm minus base, or arm minus the other arm).

		Adjusted (98.33%)		Unadjusted 95%	
Endpoint	Point	one-sided bounds	two-sided	two-sided	State
C_R (R_{21} cut, %)	23.56	[22.66, 24.46]	[22.55, 24.57]	[22.73, 24.39]	P
C_S (S_{11} cut, %)	23.39	[22.43, 24.34]	[22.31, 24.46]	[22.51, 24.27]	P
$D = C_R - C_S$	+0.17	[−0.52, +0.86]	[−0.60, +0.95]	[−0.46, +0.80]	P
MV, R_{21} − base	+0.02	[−0.51, +0.53]	[−0.57, +0.60]	[−0.46, +0.49]	P
legacy, R_{21} − base	+0.39	[−0.15, +0.93]	[−0.22, +1.00]	[−0.10, +0.89]	P
closure, R_{21} − base	+6.81	[+6.10, +7.53]	[+6.02, +7.62]	[+6.16, +7.47]	P
MV, S_{11} − base	+0.05	[−0.49, +0.60]	[−0.56, +0.66]	[−0.45, +0.55]	P
legacy, S_{11} − base	−2.40	[−3.08, −1.74]	[−3.17, −1.65]	[−3.03, −1.79]	U
closure, S_{11} − base	+7.41	[+6.65, +8.19]	[+6.56, +8.29]	[+6.71, +8.13]	P
MV, R_{21} − S_{11}	−0.04	[−0.48, +0.41]	[−0.54, +0.46]	[−0.45, +0.37]	P
legacy, R_{21} − S_{11}	+2.80	[+2.26, +3.34]	[+2.20, +3.41]	[+2.30, +3.30]	P
closure, R_{21} − S_{11}	−0.61	[−0.86, −0.37]	[−0.89, −0.33]	[−0.84, −0.38]	P
MV, S_{11} − R_{21}	+0.04	[−0.41, +0.48]	[−0.46, +0.54]	[−0.37, +0.45]	P
legacy, S_{11} − R_{21}	−2.80	[−3.34, −2.26]	[−3.41, −2.20]	[−3.30, −2.30]	U
closure, S_{11} − R_{21}	+0.61	[+0.37, +0.86]	[+0.33, +0.89]	[+0.38, +0.84]	P

number. Per problem (2,000 problems), the cut is base minus model thinking tokens, each a four-draw mean of per-draw medians. The observed per-problem cut was negative on 10.75% and 12.65% of problems for full-vector and on-axis reconstruction; these descriptive shares include sampling noise and do not establish lengthening on individual problems. Ranked on two draws and evaluated on the other two (both directions averaged), the 200 problems with the largest cuts carried 28.2% and 32.6% of the net cut (negative cuts included; an even spread would give 10%). The rank correlation of per-problem cuts between the two arms, across draws (0.530; descriptive 95% bootstrap interval 0.495 to 0.563), was of similar size to each arm's own cross-draw correlation (0.546, 0.506 to 0.583; 0.552, 0.511 to 0.591). Descriptive intervals use 2,000 problem-level bootstrap resamples (seed

Table 20: Study 3 held-out levels per arm (2,000 problems, sixteen outputs each).

Arm	Mean think tokens	MV accuracy	Legacy accuracy	Closure
Base	5,784	0.789	0.669	0.906
R_{10}	4,936	0.791	0.675	0.952
R_{14}	4,693	0.792	0.676	0.965
R_{21}	4,421	0.789	0.673	0.974
S_6	4,955	0.796	0.678	0.954
S_8	4,645	0.794	0.660	0.974
S_{11}	4,431	0.790	0.645	0.980

Table 21: Study 3 grader agreement counts per arm. Rows: outputs per arm. Agree: both graders give the same verdict. MV only, legacy only: credited by one grader alone. MV credited, unclosed: credited by Math-Verify without closing `</think>` (all), and those among them with no `\boxed` (unboxed). Unboxed, MV yes, legacy no: no `\boxed` anywhere, credited by Math-Verify and not by the legacy grader (all), and those among them that close `</think>` (closed).

					MV credited, unclosed		Unboxed, MV yes, legacy no	
Arm	Rows	Agree	MV only	Legacy only	all	unboxed	all	closed
Base	32,000	28,143	3,855	2	221	209	270	61
R_{10}	32,000	28,264	3,734	2	98	91	184	93
R_{14}	32,000	28,280	3,719	1	54	46	181	135
R_{21}	32,000	28,266	3,733	1	45	41	251	210
S_6	32,000	28,203	3,795	2	99	98	281	183
S_8	32,000	27,706	4,291	3	51	48	778	730
S_{11}	32,000	27,353	4,643	4	39	38	1,151	1,113

20260926), separate from the primary inference procedure; the cross-fit halves are draws {173, 271} and {619, 827}. Per-problem cuts are thus only moderately repeatable within an arm, and similar correlations establish neither that the arms shorten the same problems nor that their per-problem behavior is equivalent.

F Study 3 per-draw descriptive states

Study 3's frozen analysis estimates every endpoint from all four held-out draws together. For description only, we applied the same frozen estimator and state rules, unchanged, to the outputs of each draw alone (draw seeds 173, 271, 619 and 827): the same paired bootstrap and seed, the same adjusted bounds with divisor 3, and the same margins. There is no further adjustment for four draws, because these states are descriptive and are not tests. The per-draw computation was written to its own record after the four-draw readout was released; it changes no primary estimator, invalidity rule, flag, state or cell, and where it differs from the four-draw analysis the four-draw analysis governs.

The per-draw states are not independent of one another. As the per-draw reporting rule states: "with the same seed and the same n = 2000, all four draws resample the **same** problem indices. Their intervals are therefore not independent bootstrap draws, and four per-draw states are not four independent tests."

Several quantities have no per-draw definition and were not computed: a direction state for D (in Study 3, D is a matching endpoint), a resolution-split label for D and for the accuracy states, and a per-draw target-window state. A per-draw estimator refusal would have been recorded for that draw only, without affecting the four-draw result; none occurred, and all four draws returned states. Per-draw flags are reported here and are not added to the four-draw flags. Before the record was written, the same reduction reproduced every four-draw endpoint exactly, an independent reimplementation agreed with every per-draw point and bound to within 10^{-9} pp, and every per-draw state was re-derived from its bounds.

Table 22 gives the per-draw points of the four contrasts that the frozen draw-sign flag examines, and Table 23 the per-draw states. Each draw carries a quarter of every problem's outputs, and at every

target and draw the per-draw one-sided interval on D is wider than the four-draw interval. Per-draw matching is unresolved in every draw at every target, including 22.5, where the four-draw analysis establishes it; this is a descriptive property of single-draw intervals and does not qualify the four-draw state. In the names of the frozen record (for example `mv_RB` is Math-Verify accuracy of R minus the base model), over all fifteen metrics and not only the four the flag examines, the per-draw points take both signs for: `D, mv_RB, legacy_RB` (12.5); `mv_RS` (17.5); `D, mv_RB, mv_SB, mv_RS` (22.5). A per-draw state that is unresolved in some draws and resolved in others occurs for: `none` (12.5); `none` (17.5); `closure_RS` (22.5).

Table 22: Study 3 per-draw points (pp, three decimals so that signs near zero stay visible). $D = C_R - C_S$; accuracy and closure contrasts are full-vector minus on-axis. Descriptive; the draws share problems and bootstrap seed.

Target	Draw seed	D	MV, R−S	legacy, R−S	closure, R−S
12.5	173	+0.161	−0.200	−0.188	−0.362
12.5	271	+0.857	−0.413	−0.125	−0.062
12.5	619	+0.308	−0.637	−0.412	−0.038
12.5	827	−0.005	−0.750	−0.513	−0.313
17.5	173	−1.422	−0.062	+1.863	−1.075
17.5	271	−0.392	+0.150	+1.825	−0.900
17.5	619	−1.317	−0.825	+1.187	−1.112
17.5	827	−0.169	−0.013	+1.500	−0.612
22.5	173	−0.612	+0.013	+3.263	−0.500
22.5	271	+0.509	+0.400	+3.238	−0.600
22.5	619	+0.821	−0.287	+2.487	−0.612
22.5	827	−0.045	−0.275	+2.200	−0.713

Table 23: Study 3 per-draw descriptive states. Match: D within ± 1 pp (P) or unresolved (U). Accuracy (MV, legacy): E within the 2.6-pp margin, R or S resolved toward full-vector or on-axis reconstruction. Cuts: positive-cut state for R / S. Components: the six quality components in the order MV, legacy, closure against the base model, then MV, legacy, closure against the other arm; Q is their aggregate. Flags and cell are the frozen classifier's per-draw outputs. Not independent tests.

Target	Draw	Match	MV	Legacy	Cuts	R components	Q_R	S components	Q_S	Flags	Cell
12.5	173	U	E	E	P/P	PPP PPP	P	PPP PPP	P	none	MATCH-UNRESOLVED
12.5	271	U	E	E	P/P	PPP PPP	P	PPP PPP	P	none	MATCH-UNRESOLVED
12.5	619	U	E	E	P/P	PPP PPP	P	PPP PPP	P	none	MATCH-UNRESOLVED
12.5	827	U	E	E	P/P	PPP PPP	P	PPP PPP	P	none	MATCH-UNRESOLVED
17.5	173	U	E	R	P/P	PPP PPU	U	PPP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED
17.5	271	U	E	R	P/P	PPP PPU	U	PPP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED
17.5	619	U	E	R	P/P	PPP PPU	U	PPP PPP	P	GRADER-SPLIT	MATCH-UNRESOLVED
17.5	827	U	E	R	P/P	PPP PPU	U	PPP PPP	P	GRADER-SPLIT	MATCH-UNRESOLVED
22.5	173	U	E	R	P/P	PPP PPP	P	PUP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED
22.5	271	U	E	R	P/P	PPP PPU	U	PUP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED
22.5	619	U	E	R	P/P	PPP PPU	U	PUP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED
22.5	827	U	E	R	P/P	PPP PPU	U	PUP PUP	U	GRADER-SPLIT	MATCH-UNRESOLVED

G GRADER BEHAVIOR

Definitions. Every output is scored by both pinned graders, and neither is a fallback for the other. Math-Verify 0.8.0 (with SymPy 1.13.1 and antlr4-python3-runtime 4.13.2) parses the full completion, reasoning included, extracting one unanchored answer candidate (Appendix B) with first-match fallback and five-second parse and verify timeouts; it needs no box. The legacy grader requires a box: it takes the last `\boxed{}` answer, normalizes it and compares it with the gold answer as a string and then as a number, so a completion without a box is scored incorrect. Closure is the literal presence of `</think>`. Wrong, malformed, capped and unclosed outputs stay in every denominator. We report both graders and choose neither.

Disagreement categories. For each arm, Table 21 counts outputs on which the graders agree, outputs credited by Math-Verify alone (MV only) and by the legacy grader alone (legacy only). Two further categories locate the Math-Verify-only credit. "MV credited, unclosed" counts outputs credited by Math-Verify that never close `</think>`, with the unboxed ones among them (text containing no `\boxed`). "Unboxed, MV yes, legacy no" counts outputs whose text contains no `\boxed` anywhere and that Math-Verify credits and the legacy grader does not, with the closed ones among them. Legacy-only credit is rare in every arm, so almost all disagreement is Math-Verify-only credit.

An identity. Write N for the number of outputs per arm and M_a and L_a for arm a's MV-only and legacy-only counts. Every problem contributes the same sixteen outputs to every arm, so averaging within problems and then over problems equals pooling outputs, and

$$\Delta_{RS}^{\text{legacy}} - \Delta_{RS}^{\text{mv}} \;=\; \frac{100}{N}\Big[(M_S - L_S) - (M_R - L_R)\Big] \text{ pp},$$

where Δ_{RS}^{g} is the full-vector minus on-axis accuracy under grader g. The legacy-minus-Math-Verify gap is therefore exactly the on-axis arm's excess of net Math-Verify-only credit. This is arithmetic, not a finding.

A descriptive decomposition (post hoc). After the readout was released, we asked which category of Math-Verify-only credit carries the excess. Table 24 sets the gap beside the on-axis arm's excess of closed, unboxed outputs that Math-Verify credits and the legacy grader rejects. At 17.5 and 22.5 the two are of about the same size; at 12.5 both are small. This comparison was not planned before the readout, carries no interval and no test, and changes no frozen state, cell or verdict. It does not show that either grader is wrong, estimate a "true" accuracy difference, or explain why the on-axis arms produce more such outputs; within its stated limits, the model-assessed audit below describes whether these outputs state correct final answers; it is not an adjusted accuracy or contrast. Descriptively, at 17.5 and 22.5 the graders' disagreement about the objective comparison is of the size of one kind of output, a closed answer without a box, whose treatment is a grading-policy choice that the comparison inherits.

Table 24: Study 3 grader decomposition (post hoc, descriptive; pp of outputs unless counts). Gap: legacy R−S minus MV R−S accuracy difference. Net MV-only excess: on-axis minus full-vector rate of MV-only minus legacy-only outputs (equal to the gap by the identity above). Closed, unboxed excess: on-axis minus full-vector rate of closed outputs with no box that Math-Verify credits and the legacy grader does not; counts are per arm out of 32,000.

Target	Pair	Gap	Net MV-only excess	Closed, unboxed excess	Counts R / S
12.5	R_{10} / S_6	0.19	0.19	0.28	93 / 183
17.5	R_{14} / S_8	1.78	1.78	1.86	135 / 730
22.5	R_{21} / S_{11}	2.83	2.83	2.82	210 / 1,113

A model-assessed audit at 22.5 (post hoc, exploratory). After the readout was released, we audited the closed, unboxed outputs that Math-Verify credits and the legacy grader rejects at the matched target: 210 for the full-vector arm, all assessed, and 1,113 for the on-axis arm, of which a random sample of 300 was assessed. The primary analysis was fixed before any sampled answer was read; a lenient-parse sensitivity was added during the assessor run, after one masked assessor reply was inspected to diagnose parse failures, and before the analysis was run. Each output was labeled in two passes by three primary language-model assessors, requested as `claude-opus-5-5`, `claude-fable-5-1` and `claude-sonnet-5` (the served model identifier was recorded for every call and had to include the requested one). Pass 1, blind to the reference answer, saw the problem and the response after the thinking block, not the reasoning, and extracted the single final answer stated after the reasoning, if any; pass 2 saw the problem, the reference answer and that extracted string only, and judged whether they are equivalent. The primary label is the category named by at least two available readings; otherwise the output is unresolved and counted as not correct. A fourth assessor from another vendor, requested as `gpt-6-astra`, whose interface does not report the served model, enters only a sensitivity. One of the three primary assessors did not

pass its validity checks and was unavailable on many outputs, so the majority is often effectively over two assessors. By that majority, the on-axis arm has an estimated 898 more correct stated final answers among these outputs than the full-vector arm (95% interval 882 to 904 under the primary assumptions), against the 903-output difference in counts. Outputs are not paired across arms, so this is a difference between two arm totals, not a count within a fixed set of outputs, and its interval can exceed that difference. Under the worst-case assessor envelope on both arms the interval is 369 to 1,004.

Under the primary assumptions (majority labels taken as correct, an output without a majority counted as not correct, the full-vector count taken as exact, on-axis sampling as the only source of interval uncertainty, and a strict parse with one assessor unavailable on 124 of 300 sampled on-axis outputs and 98 of 210 full-vector outputs), the on-axis arm has at most 21 more outputs without a correct stated final answer than the full-vector arm at the upper 95% end. That bound does not hold under a symbolic-tool label (85) or under the worst-case envelope on both arms (534). The on-axis sample was drawn at random with a fixed seed, and its interval is a Korn-Graubard sampling interval with a problem-cluster design effect and finite-population correction, which covers sampling but not assessor error; the envelope and symbolic-tool sensitivities are defined in the audit record released with the code. Under a lenient-parse sensitivity the estimate is 905 (95% interval 895 to 905); the primary analysis governs. The audit is post hoc and exploratory, its assessors are language models whose agreement is not accuracy, and there is no human ground truth. It supports a predominantly formatting-based reading of the 22.5 grader split, subject to assessor uncertainty and parsing sensitivity, and it changes no frozen state, cell, verdict, adjusted accuracy or contrast.

Study 1. Study 1 also reports both graders. Its sealed record gives each arm's Math-Verify and legacy correct counts, and the Math-Verify-minus-legacy row rate is 10.59 pp for the base model, 11.14 pp for the full-vector finalist and 15.39 pp for the on-axis fallback. These are pooled row rates, not the frozen per-problem estimator, and they are descriptive. They are consistent with the on-axis fallback's tradeoff cell resting on its legacy safeguards (Appendix C), and they change no Study 1 state. Study 1's early-era project exposure was not certified (Appendix C).

H Execution record and operational amendments

This appendix is a dated record of how Study 3 was executed, including every change made after its design was frozen. All times are UTC in 2026. Study 1's and Study 2's provenance and scorer disclosures are in Appendices C and D.

What was frozen, and when. Study 3's design was frozen on September 22 at 02:53, before any of its new fits, development jobs or outcomes existed (its menu also reuses Study 2's fourteen adapters). The freeze fixed the two 23-alpha menus, the three targets and windows, the feasibility rule, the 1-point matching tolerance, the tie-break, the development and held-out panels, draws and seeds, both graders, the estimators, the bootstrap, the multiplicity rule, the state and cell tables, a one-replacement rule for technical failures, and the deadlines: development selection by September 23 at 11:59 and verified artifacts by September 25 at 11:59. A versioned amendment written minutes later, still before any Study 3 job, made all development generations fresh (no reuse of Study 2 development rows), which gives 376 development jobs, and fixed how held-out targets are labeled. A correction the same night states that regenerating development outputs does not remove prior knowledge of Study 2's candidates, so Study 3 is an outcome-informed follow-up, not outcome-blind. Fifteen conditions on the original development block reuse the earlier seed, configuration and order and may reproduce previously seen samples.

Development-selection deadline, amended twice during collection. Both amendments were made during data collection for operational reasons; they are not part of the original pre-specification, and the deadline was not fixed at its final value from the outset. The first (September 23, about 02:50) moved the deadline from 11:59 to 18:00 on September 23 and explicitly superseded the freeze's no-extension clause, keeping the original text, after a job-catalog registration defect, errors in operating the job scheduler and a full host disk had delayed development generation. At adoption, 194 of 376 development jobs had been accepted, no development outcome had been read, selection had never been computed on real data, and no held-out job existed. The second (about 14:30) required

all 376 development jobs before selection and a selection record committed and pushed to a remote repository strictly before September 24 at 00:00, and permitted no incomplete-development selection, extra candidate, tolerance change or outcome-dependent rescue. Development generation finished at 376 of 376 accepted jobs on September 23 at 20:47. Selection ran from 20:49 to 21:09 on the complete development record; two selector implementations returned the same selection, whose record was committed and pushed at 21:09, before the amended deadline.

Launch-scheduling gate: no-go, then an explicit override. A launch-scheduling gate, pinned on September 23 at 16:05 before the size of the selected workload was known, required the complete held-out path to fit the unchanged artifact deadline. Evaluated at 21:10 on the full selected workload of 560 held-out jobs, it returned no-go (projected overrun 0.74 h); no held-out job was admitted and the workload was not reduced to fit the gate. On September 24 a prospective operational amendment, made after selection and before any held-out generation, authorized a bounded effort to obtain more evaluation capacity and forbade any change to the selected pairs, targets, sample, draws, graders, tolerances or analysis rules, any trimming of the workload, and any use of held-out outcomes to set scope. At about 03:04, after 12 concurrent evaluators had been demonstrated, the launch of all 560 jobs was authorized, explicitly overriding the gate while preserving its failed record, retaining the September 25 11:59 verified-artifact target, and adding an operational extension to 19:00.

Held-out generation and technical failures. Held-out generation began on September 24 at 07:53 and ended on September 25 at 09:23 with all 560 logical jobs accepted (224,000 rows). On September 25 between about 02:56 and 03:25, the compute infrastructure interrupted evaluation containers in one batch. 10 logical jobs, spread over three of the four draws, stopped before writing any output row and without an error record. Each was classified as a technical failure under the frozen rule, which allows one identical replacement per logical job and retains the first valid completion; each was replaced once with an identical configuration, and every replacement was accepted. The failed attempts are preserved. No job was retried because of its result, and no seed, candidate or draw was added. Descriptive per-draw summaries (Appendix F) were specified on September 24 at about 13:45, after generation had begun and before the held-out readout. They are therefore not prospectively frozen, and they change no primary estimator, invalidity rule, flag or cell.

Analysis and independent checks. The frozen primary analysis ran on September 25 at 09:48. A second implementation, written independently of the primary analysis code, recomputed the held-out reduction, all point estimates and all eight interval endpoints for each of the 15 fixed metrics of every tested pair from the accepted outputs, using the same bootstrap generator and quantile rule; it agreed with the primary analysis to an absolute tolerance of 10^{-9} percentage points at all three targets. The frozen cell predicates were not reimplemented in this check. A supplementary derived-field check, written with access to the primary analysis code, recomputed the states from the adjusted endpoints, together with the cells, flags, counts and bookkeeping, in 360 checks with no failure. The pre-specified readout hold applied only to a disagreement between the primary and independent analyses; disagreement between the two graders is a reported result and never a hold. No hold fired, and the verified artifacts were complete on September 25 at 09:58, before the 11:59 target.

Population. Study 3 was evaluated under its own eligibility standard: its held-out panel is disjoint from documented uses under declared checks, which is not a claim that the panel is universally unexposed. The panel consists entirely of problem groups that Study 2's population rule would have excluded (earlier screening candidates and groups reserved for a replication), so the two studies' held-out panels were not cleared under one exposure standard and their results are not compared as if the populations were interchangeable.

Replication. The frozen design specified an arm-blind, schedule-only rule for deciding whether to repeat the full search with additional training seeds. It was projected in advance to admit no replication within the deadline; no replication was run, and none is claimed. The fixed-recipe refit study below refits two selected recipes and is not such a replication. A separate comparison under common forced thinking budgets was commissioned, but its scientific study was not run.

Post-hoc analyses after the Study 3 readout. None of the following changes a frozen state, cell or verdict of Studies 1 to 3. The model-assessed grader audit (Appendix G) was pre-registered on

September 25 at 15:02, before any assessor output existed; its lenient-parse sensitivity was added at 15:28, during the assessor run and before the analysis ran, and errors in the first audit report, found by an independent cross-check, were corrected in a dated erratum at 18:01 before the audit was cited here. The all-recipe development formatting analysis (Appendix N) saved its plan before any new cross-tabulation; it was proposed after the audit and the refit design were known. The per-problem analysis (Appendix E.6) fixed its specification after one held-out draw had been inspected (with a second as reference) and before it was extended to all four draws.

Refit study. Its pre-registration (Appendix M) was committed on September 25 at 19:17, after the Study 3 results and the grader audit were known and before any refit existed. At 22:08:47 its pinned capacity rule chose one new training seed. All 4 fits and 83 evaluation jobs were accepted by September 26 at 01:56; one technical failure was replaced once under the pinned rule. The analysis read a listing of accepted jobs taken at 02:08:45, after all 83 jobs had been accepted and before the pinned stop time; no further job existed, so the listing equals the complete scheduled set. Two independent implementations agreed on every reported value.

I Direction construction and the causal knob

This appendix reports the construction of the halt direction, the evidence that it is a causal stopping knob, and the controls on its use in training. All results here are historical 7B results unless a tier is named; they predate Studies 1 to 3 and use their own estimators and populations. The forced-answer point, the construction of both directions, the steering sweep (Table 25) and the direction-specificity control (Table 27) were reported by Jayabahu & Adeleke (2026); the removal, end-of-think and reader-circuit diagnostics are our own measurements on the same model, using the published construction and recipe.

The forced-answer point. The stopping label follows Jayabahu & Adeleke (2026). A first-token probe of the gold answer can fire on the expected answer format before any reasoning has happened, so it is an unsafe place to cut. The forced-answer measurement truncates the reasoning at a candidate point, appends a short scaffold (`</think>` plus a boxed-answer cue), has the model emit an answer greedily, and grades that answer; B_{end} is the first prefix whose forced answer is correct. It is an operational candidate cut point under this scaffold, not a guarantee for later prefixes or for free continuation. A problem is *compressible* if much reasoning follows B_{end} and an *anchor* if none does. Because it is measured offline from the gold answer, B_{end} can label training data but cannot be used at deployment.

Construction. Following Jiang et al. (2026), Jayabahu & Adeleke (2026) construct difference-of-means directions. For a contrast set C in which each item has "before" positions T_{pre} and "after" positions T_{post}, the per-layer direction is

$$a^{(\ell)} = \frac{1}{|C|} \sum_{c \in C} \left(\frac{1}{|T^c_{\text{post}}|} \sum_{t \in T^c_{\text{post}}} h^{(\ell)}_{c,t} - \frac{1}{|T^c_{\text{pre}}|} \sum_{t \in T^c_{\text{pre}}} h^{(\ell)}_{c,t} \right), \qquad \hat{u}^{(\ell)} = a^{(\ell)} / \|a^{(\ell)}\|, \tag{2}$$

read out by cosine projection. They build two: a replication of the value axis of Jiang et al. (2026) from an in-context guessing-game contrast, used as a control, and the **halt direction**, which compares think-token positions after the gold answer's first-token probability rises to 0.9 and remains there with earlier positions, in correctly answered rollouts. Two disclosures apply. First, the halt direction contrasts positions *within* a rollout; a variant contrasting correct with incorrect rollouts was built and discarded as surface-confounded (it separated near 0.96 but was flat from layer 0). Second, the first-token probability is the measure of knowing that the forced-answer point rejects as unsafe to cut on: it places the contrast, it is never a training target, and cut points are forced-answer graded throughout. We therefore quote no held-out separation for either direction and rest the halt direction's value on the causal tests below. The two directions are near-orthogonal. Neither measures answerability reliably: the projection on the halt direction continues to rise long after B_{end}. What distinguishes the halt direction is actuation rather than separability, a claim about deployed behavior and not about the read-out.

Steering. The steering results in this paragraph and in Table 25 are those of Jayabahu & Adeleke (2026). Steering adds $(\alpha/100)\,\|\bar{h}\|\,\hat{u}$ to the residual stream at layer ℓ during generation, where $\|\bar{h}\|$ is the mean activation norm, and α is swept. On a 128-problem demonstration at layer 6, length falls as strength rises, while a magnitude-matched value-axis control barely changes length ($\rho = +0.09$). Among layers 6, 12, 18 and 24, steering at layer 18 moves length most strongly ($\rho = -0.82$; at $+10$ generation is 31% shorter and terminates more often); layer 24 keeps the most accuracy but moves length little (-0.24; Table 25). The published 7B recipe, and every held-out finalist here, therefore internalizes at layer 18; for the other model sizes, each model's best layer is a reported result rather than an assumption. Correctness under the forced-answer scaffold does not guarantee correctness after steering. Two inference-time diagnostics (Tables 25 and 26) ran at shorter caps (8k and 12k) than the 16k-token cap used for every trained arm. They are read for the sign and ordering of a steering effect, never as lengths or accuracies comparable to the 16k arms; their base rows share the treated rows' cap, so each comparison is internally paired.

Table 25: **Layer sweep for halt-direction steering**, from Jayabahu & Adeleke (2026) (64 problems, strength -50 to $+50$, **8k cap** as its cells indicate; at strength 0 the base model has 4,746 median think tokens, 80% termination and 72% accuracy). Layer 18 moves length most strongly. An inference-time diagnostic read for sign and ordering only; its absolute values are not comparable to 16k-cap results.

layer	ρ(strength, length)	$+10$: think / term / acc	$+25$: think / term / acc
6	-0.63	981 / 77% / 14%	27 / 67% / 0%
12	-0.44	3685 / 75% / 28%	8192 / 45% / 2%
18	$\mathbf{-0.82}$	**3256 / 83% / 56%**	18 / 89% / 12%
24	-0.24	4619 / 78% / 72%	3180 / 80% / 61%

Removal, end-of-think distinctness and the reader circuit. That the direction is a causal actuator does not make it the model's endogenous halting representation rather than one of several directions that shorten generation. *Removal:* projecting the layer-18 halt component out at every position, paired against the identical hook-off run on 128 problems, lengthens generation by about 1,480 think tokens, collapses termination by 23 pp with a threefold rise in backtracking, and lowers accuracy, while the same projection of a magnitude-matched random direction or of the value axis does the opposite on every axis (Table 26). With the addition result above, the halt component moves termination in both directions, direction-specifically. We do not call it necessary and sufficient: removal perturbs the component at every position and could disturb correlated computation, addition at any compressing strength also costs accuracy, and two controls establish specificity against those two directions, not uniqueness among all possible halting features; a generic positional-progress alternative remains untested. *End-of-think distinctness:* the direction that maximally raises the `</think>`/EOS logit, essentially the lever of Sheng et al. (2025), is a candidate confound, and the halt direction is not it: their cosine overlap at layer 18 is 0.004, indistinguishable from a random direction, and never exceeds about 0.05 across all 28 layers. Read through the unembedding, the halt direction leaves `</think>` near rank 39,700 of 152,064 and promotes answer-commitment words. Under matched steering the end-of-think direction is a monotone saturating terminator, whereas the halt direction is non-monotone, with termination peaking near strength 25 and then reversing. *Reader circuit:* injecting at layer 18 and patching each late-layer component back to its clean value, the effect on the `</think>` logit is about 96% reader-mediated, the direct residual-to-unembedding path carrying about 4%, and concentrated in the final two MLP layers. These are point estimates over 98 prefixes passing a 0.5-logit effect screen, consistent in direction across all eight shards, with modest magnitudes at a single injected position.

Raw steering is not the trained recipe. Under raw steering, accuracy collapses for both the halt and end-of-think directions at positive strength, to about 7% and 11% respectively at strength 25, so accuracy-preserving per-problem compression is not a property of raw steering. As a hooked-in actuator the halt direction is blunt: the coefficient large enough to compress collapses accuracy from 0.56 to 0.07. Why the halt direction's termination reverses at higher strength was not measured.

Direction specificity in training. If any layer-18 target reconstructed at B_{end} installed a halt, the effect would come from the timing label and not from the direction. Jayabahu & Adeleke (2026) test this by re-running the full-vector reconstruction recipe with only $\hat{u}$ changed, at fixed injected

Table 26: **Removing the halt component** (128 compressible problems, greedy, **12k cap**, paired against the hook-off baseline: 7,405 think tokens, 57.8% terminated, 56.3% accuracy). The low base termination is a property of this diagnostic regime (greedy decoding under a 12k cap), not of the deployed model. Every row is paired against the same baseline; the absolute values are not comparable to 16k-cap results. Where the base fails to terminate on 42% of traces, anything that raises termination converts capped, zero-scoring traces into gradeable ones, so the controls' accuracy gains are not evidence that a random perturbation improves the model; the table rests on the *sign* of the halt direction's effect against the controls'. Removing the halt component lengthens generation (paired Δthink $+1479$, 95% CI $[+658, +2325]$) and lowers termination (Δterm -0.23, 95% CI $[-0.33, -0.14]$).

project out (layer 18, every position)	Δthink	Δterm (pp)	Δacc (pp)	backtracks
halt direction	**+1479**	**−23**	−20	×3.1
random direction (control)	−715	+11	+13	flat
value axis (control)	−785	+16	+11	flat

magnitude $(\alpha/100)\|\bar{h}\|$ (Table 27). Only the halt direction compresses: on AMC it cuts 24%, against -2% for the value axis and $+3$% for an isotropic random direction, with no intervals reported. The published direction is built from 457 of a probe pool of 664 rollouts on 83 problems drawn from an offline pool; the preprint does not say which problems, so we cannot rule out that the AMC reads fall on the direction's construction problems. No adapter parameter is reported as selected on AMC.

Table 27: **Direction specificity at fixed magnitude**, from Jayabahu & Adeleke (2026). The 7B full-vector reconstruction recipe re-run with only the direction $\hat{u}$ changed. No intervals are reported.

	train-24			AMC		
direction reconstructed at B_{end}	acc	think	Δ	acc	think	Δ
halt direction	0.792	4911	−25%	0.822	4760	−24%
value axis (control)	0.771	6253	−4%	0.797	6162	−2%
random direction (control)	0.771	7195	+10%	0.807	6468	+3%

J HISTORICAL INTERNALIZATION RESULTS

This appendix collects the historical fixed-recipe evaluations summarized in Section 4. They predate Studies 1 to 3, use their own estimators, rollout counts and a 100-ID holdout, and select or fix configurations as described below. Their levels are not comparable with the held-out results of Studies 1 to 3, which use different populations and estimators. The 7B fit and its generalization, termination and seed results, the 7B scalar-projection arms, the other training objectives and the DEER baseline were reported by Jayabahu & Adeleke (2026); Tables 28 and 29 reproduce tables of that paper, and its training-set ablations, other objectives and adapter-averaging results are summarized or cited rather than reproduced. The model-size and family ladder, the 14B GPQA-Diamond evaluation, the matched-capacity scalar ladder, the DLER reads and the composition with DLER are our own runs of that recipe on public models and benchmarks.

Conventions. Accuracy is Math-Verify unless the legacy boxed-answer grader is named. Clustered paired reductions use per-problem median lengths against their own stated base; one-stage intervals resample problems and two-stage intervals also resample their rollouts. Mean think-token columns and marginal percentage changes in Tables 29 and 33 summarize the displayed arm and base means; they are not the paired median-based estimates, and the two must not be converted into each other from printed aggregates. All rows use a 16k-token cap except the inference-time diagnostics of Appendix I. A tier's *champion* is its selected full-vector reconstruction adapter. *Historical "held accuracy"* is the operational rule of Jayabahu & Adeleke (2026), not the quality rule of Studies 1 to 3: accuracy within about 2 pp of base (roughly its 2.6 pp cross-seed standard deviation over three seeds), with termination no worse than the base model, read as a point comparison rather than a resolved measurement. Non-resolution is not equivalence.

J.1 THE 7B FIT

Jayabahu & Adeleke (2026) fit a low-rank adapter to 24 compressible problems without reinforcement learning, against a base at 0.812 accuracy, 6,516 think tokens and 0.95 termination. Full-vector reconstruction at $\alpha = 25$ cut think tokens 25% at 0.792 accuracy, the only listed training-set arm within their historical held-accuracy rule; imitation and steering-distillation arms were longer or less accurate. The strength was chosen as the strongest setting with accuracy within the 2.6-point threshold on those 24 problems, where such differences cannot be resolved, and their schedule, strength and capacity grid is not reproduced here.

J.2 GENERALIZATION AT 7B

Jayabahu & Adeleke (2026) evaluated the 24-problem adapter zero-shot at a 16k cap on a 100-problem compressible holdout and four external benchmarks (Table 28). It removed 20 to 24% of thinking on all five, including MATH500 at about 3,100 think tokens, with accuracy changes from -2.0 to $+2.7$ points; no intervals are reported. The largest accuracy change, $+2.7$ points on AIME25, accompanies the largest closure gain, 0.62 to 0.91: traces previously capped and scored zero now terminate and become gradeable. None of this establishes held accuracy across the board: in our runs at 1.5B and 14B the AIME25 cost resolves below zero (Appendix J.4).

Table 28: **Held-out generalization at 7B**, from Jayabahu & Adeleke (2026): the 24-problem adapter zero-shot at a 16k cap. The holdout, described by Jayabahu & Adeleke (2026) as 100 compressible problems disjoint from the training set, is not certified separate from all development data (Appendix J.3); its row is graded at $n = 4$, and the matched $n = 8$ read is 0.819 accuracy at -22.0%. No intervals are reported.

eval set	base (acc / think)	recon (acc / think)	Δthink	closed
compressible holdout	0.845 / 6617	0.835 / 5014	−24%	0.98
AMC	0.813 / 6280	0.822 / 4760	−24%	0.98
AIME24	0.512 / 10413	0.492 / 7898	−24%	0.94
AIME25	0.358 / 10808	0.385 / 8376	−22%	0.91
MATH500 (Hendrycks et al., 2021)	0.934 / 3101	0.930 / 2495	−20%	0.99

Association with per-problem slack. Jayabahu & Adeleke (2026) correlate each holdout problem's realized think cut with its own compressible slack, a reference the model never sees, and report a Pearson correlation of 0.70 (95% CI 0.25 to 0.85) at strength 25. In our runs the absolute-token correlation also resolves at 14B, while the 1.5B ratio-form interval spans zero (Table 30). This does not establish an advantage over a global length policy.

Termination, seeds and a science benchmark. As Jayabahu & Adeleke (2026) report, the base model more often fails to terminate within the 16k cap on harder benchmarks, and the adapter reduces non-termination on every benchmark, on the two AIME sets from 31 and 38% to 6 and 9%. Averaged over three seeds, the cut is $25.0\% \pm 2.3$ points on the training problems and $25.1\% \pm 1.0$ on AMC (Table 29). In our runs, the 14B champion cuts 18.3% (pooled) on GPQA-Diamond (Rein et al., 2024) zero-shot without establishing held accuracy: its paired accuracy change is -1.4 points $[-4.9, +2.0]$ over 198 problems, wider than the 2.6-point threshold. The 1.5B tier is excluded because its base accuracy, 0.307, is below the 0.31 minimum fixed in advance for an interpretable read on a four-option benchmark.

Table 29: **Per-seed detail** (7B), from Jayabahu & Adeleke (2026). Seed 17 is the reported adapter; closure stays at or above 0.97 for every seed (not tabulated here).

n/a	seed 17	seed 0	seed 42	mean ± sd
train-24				
Δthink	−25%	−23%	−28%	−25.0% ± 2.3
acc	0.792	0.771	0.823	0.795 ± 0.026
AMC				
Δthink	−24%	−25%	−26%	−25.1% ± 1.0
acc	0.822	0.818	0.825	0.822 ± 0.004

J.3 Data provenance

The published 7B halt direction is built from 457 of a probe pool of 664 base-model rollouts on 83 problems (Jayabahu & Adeleke, 2026); which problems is not stated, so we cannot rule out that the AMC columns' specificity and seed controls are read on the direction's construction problems. No adapter parameter is reported as selected on AMC. The development data behind that recipe (direction construction, B_{end} scans, the layer sweep, the steer scan and the recipe grid) are not fully described, so separation of the 7B holdout from all development data is not established.

J.4 Model sizes and a second family

We re-ran the published recipe end to end on R1-Distill-Qwen-1.5B, R1-Distill-Qwen-14B and R1-Distill-Llama-8B, each model deriving its own direction, B_{end}, layer, sign and strength (Table 30). The halt direction exists at every Qwen size with the same halt-promoting sign and no usable raw-steering operating point; its causal band tracks depth, not a fixed index. Reconstruction at $\alpha = +25$, untuned, cuts 33.1% at 1.5B and 23.2% at 14B on each tier's own holdout (24% at 7B as published). Unlike the 7B, both new tiers selected their champion layer and seed by holdout think cut at held accuracy (14B: layers 24 and 27 across seeds 0, 17 and 42, resolved to layer 24 and seed 17; 1.5B: layers 14 and 18, resolved to layer 14 and seed 17), so those holdout cells are selected quantities, not held-out evidence. The scaling evidence is the four external benchmarks, never selection sets at any tier, with cuts of 16.94 to 30.41% (Table 31). Those surfaces are disjoint from the selection population but read off the same selected layer and seed, so they show that the cut is not a property of the holdout problems, not that it is independent of the checkpoint choice. Neither new tier compresses accuracy-free everywhere. At 14B, AIME25 resolves a cost (−4.2 points $[-8.5, -0.6]$); at 1.5B, AIME25 (−3.5 $[-7.1, -0.4]$) and MATH500 (−2.1 $[-3.4, -0.7]$) resolve, while the 1.5B holdout and AMC cells are grader-split. Magnitude tracks difficulty, while resolution also tracks sample size (MATH500 has 500 problems against AIME's 30). Restated over each tier's base-terminating problems only, the median cut is −30.8% $[-35.8, -25.9]$ at 1.5B and −21.2% $[-25.4, -15.8]$ at 14B, so the pooled cells are mildly inflated by rescued traces (traces capped under the base model that now terminate).

Direction specificity at the new tiers. Re-running each tier's champion with only the reconstructed direction swapped reproduces the 7B control, but the controls fail differently. At 14B the halt direction gives a paired median of −1,266 think tokens at +1.8 points accuracy against −124 for a magnitude-matched random direction at +2.4 points; at 1.5B the guessing-game direction does not compress (paired median −52) and the random direction shortens only by losing accuracy (−486 at −5.9 points). At both new Qwen tiers the champion is the only arm that compresses at held accuracy, though on the 1.5B that verdict is grader-split. At 14B, a length-loaded difficulty direction, reconstructed through the same recipe at the same strength, moves length by 2.4% against the halt direction's 22.6% (both marginal reads from the same paired run against a shared base). A pre-registered steering comparison of the two directions discriminated nothing, because at its strength neither arm halted; that is a null about a steering probe, not evidence about the installed halt.

Llama-8B and Phi-4. On R1-Distill-Llama-8B the compressible gap survives (the forced-answer point sits at 0.51 of the answer-reaching length), and the halt direction exists, but its sign inverts: subtracting it at a deep layer ($\alpha = -25$ at layers 24, 26 and 28) is what shortens. Its raw-steering cut

Table 30: **Family-and-size ladder.** Each tier uses its own direction and recipe. Qwen holdout cuts are pooled reductions of per-problem medians; adaptivity is the absolute-token correlation. The 1.5B and 14B rows are our runs; their champions were selected on their holdouts. [§]Published by Jayabahu & Adeleke (2026) (marginal holdout cut; its slack correlation is reported in ratio form, 0.70). [*]Intervals: $+0.62$ $[+0.44, +0.76]$ at 1.5B, $+0.65$ $[+0.52, +0.75]$ at 14B; the ratio form's 1.5B interval $[-0.06, +0.76]$ includes zero. [†]Llama-8B's correlation is positive: subtracting the direction ($\alpha = -25$ at layers 24, 26 and 28) is what shortens. [‡]Reconstruction null at layer 28, the one layer clearing the pre-registered steer-hygiene criterion: $+148.9$ median think tokens, 95% CI $[-196.0, +338.5]$.

model	pool	band	steer ρ	recon ℓ	holdout cut	adapt. r
Qwen-1.5B	1008	14–18	-0.71	14	-33.1%	$+0.62^{*}$
Qwen-7B[§]	n/a	n/a	-0.82	18	-24%	n/a
Qwen-14B	536	24–27	-0.90	24	-23.2%	$+0.65$
cross-family						
Llama-8B	929	24, 26, 28	$+0.33^{\dagger}$	n/a	null[‡]	unmeasured

Table 31: **Cross-model benchmark cells.** Paired base versus each tier's champion, 16k cap, eight rollouts per problem (16 on AIME). *Equivalent*: 90% paired clustered interval inside ± 2.6 points; *inconclusive*: neither equivalence nor a resolved cost; *unresolved*: negative point estimate with no clustered interval available; **bold**: interval entirely below zero. [‡]Grader-split: on the holdout the legacy grader reads -8.8 $[-14.1, -3.9]$, on AMC -5.4 $[-9.6, -1.5]$. [§]Resolved under both graders; the legacy grader reads -6.4 $[-8.3, -4.6]$. [*]Full 500-problem set.

benchmark	pooled Δthink	base acc	Δacc (pp)	closure
Qwen-1.5B, champion layer 14, seed 17				
compressible holdout	-33.1%	0.766	$+0.6$ $[-3.0, +4.0]$ (grader-split)[‡]	0.97 to 0.99
MATH500[*]	-25.6%	0.830	$-$**2.1** $[-3.4, -0.7]$[§]	0.94 to 0.98
AMC	-29.6%	0.658	-2.6 $[-6.3, +1.1]$ (grader-split)[‡]	0.83 to 0.95
AIME24	-30.4%	0.292	-2.3 (unresolved)	0.64 to 0.92
AIME25	-30.0%	0.240	$-$**3.5** $[-7.1, -0.4]$	0.68 to 0.93
Qwen-14B, champion layer 24, seed 17				
compressible holdout	-23.2%	0.868	$+1.8$ $[-1.2, +4.8]$ (inconclusive)	0.99 to 0.99
MATH500[*]	-19.1%	0.946	-0.2 (unresolved)	0.99 to 0.99
AMC	-18.3%	0.863	$+0.0$ $[-2.0, +2.1]$ (equivalent)	0.94 to 0.98
AIME24	-20.0%	0.660	-2.1 (unresolved)	0.83 to 0.94
AIME25	-16.9%	0.485	$-$**4.2** $[-8.5, -0.6]$	0.74 to 0.93

is marginal and layer-attributed, 2.6% in median and 5.4% in mean at layer 24, and the axis-swap control does not establish specificity there. Reconstruction at that model's own sign gives a null at the one layer clearing the pre-registered steer-hygiene criterion (+148.9 median think tokens, 95% CI $[-196.0, +338.5]$); the band's other four layers remain unreadable under that criterion. On Phi-4 the pipeline did not meet its pre-registered feasibility gate (the forced-answer instrument locked on for 83.5% of problems against a 90% threshold), so no reconstruction was run; this is a statement about the pipeline, not about the model, whose median compressible slack is 0.711.

J.5 Historical scalar-projection comparisons

These arms trained a scalar-projection (hinge) target, which rewards the projection of the layer-18 activation onto $\hat{u}$ up to a margin and leaves off-axis coordinates unconstrained. They are not on-axis reconstruction, the objective of Studies 2 and 3 (Equation 1). With added MLP capacity or a raw (unnormalized) projection, Jayabahu & Adeleke (2026) found that the tested variants grew longer and less accurate as their targets increased; the strongest arm raised the cosine projection to +0.34 while its generations grew 43% longer, its accuracy dropped to 0.57 and its closure dropped from 0.95 to 0.61.

Matched capacity at 1.5B and 14B. At each tier's own layer, sign, seed and 24 problems, attention-only, with three hinge margins bracketing the published 7B sweep, we compared scalar arms against the full-vector reconstruction champion (Figure 3). Reconstruction's layer and seed were selected on the displayed holdout and the scalar arms were then evaluated at that configuration: a fixed-recipe ablation conditional on a reconstruction winner, not a comparison after equal selection. At 1.5B the smallest margin shows no resolved compression (-5.7%, $[-13.7, +2.3]$) and the two larger ones compress only at a resolved accuracy cost (-11.8% at -8.4 points $[-11.6, -5.2]$; -16.0% at -6.6 $[-10.5, -2.7]$). At 14B the pattern repeats: -4.1% spanning zero, then -6.7% at -6.8 points and -12.6% at -11.1. All six scalar arms close fewer traces than their own base, with one-stage clustered intervals from -2.1 $[-4.0, -0.5]$ to -6.8 $[-9.5, -4.3]$; of the two champion gains only the 1.5B's resolves ($+2.0$ $[+0.6, +3.5]$). All six scalar arms moved on-axis, but the available check of on-axis movement reports medians rather than the pre-registered clustered intervals.

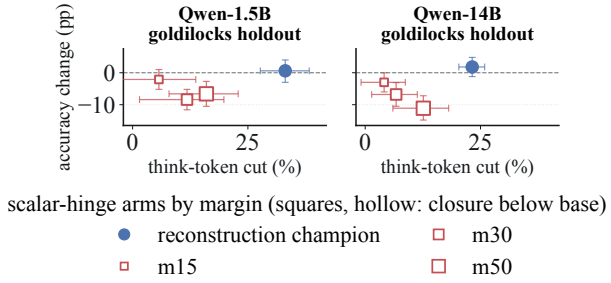

Figure 3: **Fixed-recipe comparison of full-vector reconstruction and scalar-hinge targets at 1.5B and 14B** (our runs). Think-token cut against paired accuracy change, reconstruction champion (circles) against matched-capacity scalar-hinge arms (squares, sized by hinge margin), with clustered 95% intervals. Marker fill encodes the termination half of the historical held-accuracy rule (filled: closure at least the arm's own base). Populations are per panel and not pooled: each panel is that tier's own holdout, on which its champion was selected. Reconstruction's configuration was selected first and imposed on the scalar arms. No admissible region is drawn. Panel titles say "goldilocks holdout" for the compressible holdout.

J.6 Other training objectives

Jayabahu & Adeleke (2026) report that outcome reinforcement learning with per-problem budgets, imitation and preference optimization did not install a per-problem halt as implemented there; these are negatives about those implementations, not general impossibility claims, and a heavily optimized length-penalty pipeline does compress (Appendix J.7). One caution for reproduction: the DeepSeek-

R1 chat template silently strips <think> and </think> from assistant messages, so training text must be composed manually.

J.7 Baselines and composition

Baselines. Two baselines compress harder (Table 32). The reimplementation of the decoding-time early exit DEER (Yang et al., 2025a) by Jayabahu & Adeleke (2026) cuts more but costs up to eight accuracy points on the hard sets; it runs greedily at $n = 1$ against a matched never-exit anchor, so it is placed on the same axes as the reconstruction arms, not matched to them decode for decode. In our evaluation, DLER (Liu et al., 2025), a length-penalty reinforcement-learning model fine-tuned from the same base, cuts about 66% while gaining +1.4 points under Math-Verify, averaged over the four external benchmarks, out-compressing this recipe roughly threefold; its accuracy is grader-split. We do not contest the ranking.

Table 32: **Baselines, each method against its own base.** Signed think-token change (%) and accuracy change (pp). The reconstruction and DEER rows are from Jayabahu & Adeleke (2026) (reconstruction sampled with $n \geq 4$; DEER greedy with one rollout against a never-exit anchor); the DLER row is our evaluation of the released DLER model. ‡DLER's MATH500 cell uses a 484-problem clean subset, versus reconstruction's 500, so the two are not comparable.

n/a	Δthink % / Δacc (pp), each method vs its own base				
method	holdout	AMC	AIME24	AIME25	MATH500
full-vector reconstruction	−24 / −1.0	−24 / +0.9	−24 / −2.0	−22 / +2.7	−20 / −0.4
DLER (Liu et al., 2025)	n/a	−65 / +4.2	−71 / −0.2	−73 / +0.6	−53.4 / +1.1‡
DEER (Yang et al., 2025a)	−64 / +5.0	−40 / −8.4	−26 / −6.7	n/a	−54 / −4.0

Composition with length-penalty training. We re-ran the entire published recipe with DLER as the base (direction, B_{end} and the 24 problems re-derived on DLER; layer 18, strength 25, hold schedule, attention-only, seed 17). It removes a further 11.7 to 14.9% of thinking on all four external benchmarks, with closure 1.00 in every cell (Table 33). This is not a held-accuracy result: three of the four resolve accuracy costs from 1.3 to 7.1 points, and AMC's interval spans zero. It is a single-seed demonstration without a random-direction control on DLER. The direction remains a causal length lever on DLER's weights ($\rho = -0.90$ at layer 18 over 128 problems).

Table 33: **Additional shortening after DLER.** Mean-token marginal reductions against DLER alone; paired accuracy intervals use problem clusters; **bold** marks an interval entirely below zero. MATH500 accuracy uses the 484-problem clean subset.

arm (mean think tokens)	AIME24	AIME25	AMC	MATH500
DLER alone (Liu et al., 2025)	2886	2876	1881	1174
full-vector reconstruction on DLER	2457	2449	1634	1037
Δthink	−14.9%	−14.8%	−13.1%	−11.7%
closure (both arms)	1.00	1.00	1.00	1.00
Δacc (pp)	**−7.1**	**−2.5**	−0.9	**−1.3**
95% CI	[−11.2, −3.3]	[−4.8, −0.4]	[−3.0, +1.1]	[−2.0, −0.7]

K Attempted mechanism study and a second direction

This appendix reports what we measured about why training objectives might differ, and a test of whether the historical objective comparison generalizes beyond halting. None of it supplies a mechanism result. Off-axis drift was measured only on historical arms; it was not measured in Studies 1 to 3.

K.1 HISTORICAL OFF-AXIS DRIFT

Teacher-forced drift was measured on the historical arms (at 7B, as reported by Jayabahu & Adeleke (2026)): adapter-on against adapter-off activations at layer 18 on think tokens, with on-axis movement normalized by activation norm and off-axis drift by injected magnitude. Greater normalized off-axis drift accompanies the scalar-hinge outcomes. At 7B, raw drift and drift per unit of achieved on-axis movement both separate the tested families; mass beyond the base model's own 99th percentile is 0.0% under full-vector reconstruction and at least 0.2% for every scalar arm. In our runs of the recipe at 1.5B and 14B the normalized ordering repeats and the raw one does not (the 1.5B reconstruction arm's raw drift of 0.352 slightly exceeds one scalar arm's 0.347). These observations motivated a reader-disruption hypothesis, but the refusal experiment below does not identify drift as the cause of an objective's behavioral cost. For weight-space editing generally, drift is a marker worth tracking; it is not a validated safety criterion. The same work reports that across its ten arms higher drift goes with lower training-set accuracy (midrank Spearman -0.80, exact $p = 0.008$; 0.03 after a fourfold correction). We draw no admissible region from drift. The drift measurements compare trained adapters with base, so they do not attribute raw-steering degradation to drift or establish drift as necessary for damage.

K.2 THE ATTEMPTED SELECTIVE-RESTORATION STUDY

A separate study was designed to ask, with fresh matched arms, whether restoring off-axis components rescues behavior while the on-axis signal and the amount of shortening are controlled. The intervention was not validated and the behavioral study was not run; we report no mechanism result from it, and instrument failure supplies no evidence against the hypothesis. Three pilots and a successor protocol failed validation: the first two pilots' checks were insensitive to an incorrect intervention, the third failed numerical conformance at the reader boundary under a tolerance frozen before that pilot but informed by earlier diagnostics, and the successor protocol passed all 17 declared positive CPU cases and all five required mutation classes under its minimum scope, while the broader original certificate passed 21 of 23 cases. Its freeze is not certified as predating the initial CPU results, so those results are engineering evidence, not prospective confirmation, and it failed its frozen cache-equivalence check on the production model at the first declared case. No behavioral inventory or rescue, specificity or recovery contrast exists. The planned population was drawn fresh and does not bear on earlier evaluations. A follow-up would need a frozen protocol, validated cached-versus-uncached execution on the production model, discriminating controls and a passing pilot; even then it could support only the registered intervention claim, not reconstruction necessity.

K.3 A SECOND DIRECTION: REFUSAL

We pre-registered whether the full-vector versus scalar-projection dissociation reported for halting reproduces on a different direction, with the outcome table and thresholds committed before either arm was trained. It does not: both objectives install refusal, which bounds any generalization of the historical halting comparison.

The direction and causal check. A refusal direction was built as a difference of means over 100 behavior-matched harmful/harmless instruction pairs, selecting the layer by held-out AUROC; layer 22 gives AUROC 0.9348. Its absolute cosine with the halt direction is 0.0102 at that layer, against a 0.0327 chance 95th percentile from magnitude-matched random directions. Steering along it at layer 22 raises the benign refusal rate monotonically in strength (rise 0.4200 over all rollouts), while a magnitude-matched random direction produces 0.0000 refusals at every strength; layer 18 failed a pre-registered degeneracy criterion and was not carried forward. That run generated at most 256 tokens per rollout and only 85 of 400 rollouts closed the reasoning block, so the magnitude is confounded by truncation; the random direction refused nothing under the same truncation, but the confound is not discharged.

The two arms. The direction was internalized twice at layer 22 under matched conditions, once with the full-vector reconstruction target and once with the scalar-projection target, on 77 induced examples. The pre-registered primary endpoint is the change in benign refusal rate on a fresh held-out slice of 200 Alpaca instructions, against a base rate committed before either arm's completions

existed (Table 34). Both intervals lie above zero, so both targets install the behavior. The arm-to-arm difference, +4.50 pp $[-2.02, +11.02]$, contains zero, so neither arm is reported as stronger. Charging every degenerate output against its own arm as a spurious refusal still leaves both installing (+82.00 and +85.00 pp), and both arms close their reasoning on 0.9750 of rollouts. On refusal, the scalar arm's off-axis drift per unit of on-axis movement was $1.58\times$ the reconstruction arm's, the same ordering as halting, yet both installed; drift ordering did not predict installation failure on this direction.

Table 34: **Refusal installation** (layer 22, 200 held-out benign instructions). Intervals are unpaired 95% intervals on the change against base.

arm	benign refusal rate	Δ vs base	95% CI
base (unsteered)	$3/200 = 0.0150$	n/a	n/a
full-vector reconstruction	$170/200 = 0.8500$	+83.50 pp	$[+78.27, +88.73]$
scalar projection	$179/200 = 0.8950$	+88.00 pp	$[+83.43, +92.57]$

Capability. Six evaluations of one unchanged arm span 14.58 pp on the 24-problem set, because the adapter evaluation path is not reproducible run to run, so a single-draw capability cell cannot separate retention from loss. A pre-registered larger measurement averaged six evaluations per arm on the 24-problem set: full-vector reconstruction 0.7899 and scalar projection 0.8056 against a 0.8125 base, an arm-to-arm difference of $+1.56$ pp $[-2.08, +5.21]$. No arm-to-arm accuracy difference is resolved, which does not establish either arm's equivalence to base. The on-task grading leg is unreadable for either arm, because an arm refusing 85 to 90% of benign instructions leaves too few attempted answers for the grader's must-fail control. The historical full-vector versus scalar-projection distinction for halting did not replicate on this near-orthogonal direction. This is not a finding that reconstruction fails, since both arms installed.

L Additional related work and domain screen

Output-mediated and dynamic steering. Blank et al. (2026) study steering-vector transfer through fine-tuning on a steered teacher's outputs. That output-mediated route differs from direct activation-target losses; the negative imitation implementations of Appendix J are not a general impossibility claim. Shi et al. (2026) use context-specific latent steering during reasoning, a different sense of internalization from a checkpoint deployed without an intervention hook.

What LoRRA already contains. Our full-vector reconstruction objective is the LoRRA objective of Zou et al. (2023), and the overlap goes beyond the loss. That work already merges the adapter at no inference cost, already separates the layers it edits from the layer it targets, and already masks positions outside the region of interest; the writer-reader split and position masking used here follow it. It does not compare the full-vector target with an on-axis target at matched shortening, which is what Studies 2 and 3 attempt. Fixed-recipe comparisons (Jayabahu & Adeleke, 2026) favored reconstruction over hinge-type scalar targets under their operational rule, at the tested settings (Appendix J).

Relation to the halting preprint. Jayabahu & Adeleke (2026) report the halt direction, a hook-free fit on 24 problems with its cross-benchmark evaluation, and a one-model comparison with a scalar target, which they explain by off-axis drift. The three prospective held-out comparisons (Studies 1 to 3), the matched-shortening protocol and the grader analysis are new here. Our evaluations at other model sizes and families and our further diagnostics of the direction (Appendices I and J) apply its published recipe beyond that one model. We do not adopt the off-axis explanation (Section 7).

Model editing, task and weight arithmetic. Subtracting a base model from its fine-tuned version gives a task vector that can be composed with others (Ilharco et al., 2023), edits are merged across tasks (Wortsman et al., 2022; Yadav et al., 2023), and low-rank adapters (Hu et al., 2022) localize the change; Fierro & Roger (2026) isolate a behavior direction in weight space by subtracting two fine-tunes, without reading it from activations. The nearest trainable-steering relatives fine-tune adapters to produce steered *activations* rather than steered *outputs*. The representation-engineering taxonomy of Wehner et al. (2025) places our objectives in this "tune toward desired activations"

family. Within it, full-vector reconstruction also constrains the components orthogonal to the steering direction, while on-axis reconstruction constrains only the projection onto it.

Reading reasoning state. Linear probes on hidden states predict, before a chain of thought ends, whether it will succeed (Afzal et al., 2025) and whether an intermediate answer is already right (Zhang et al., 2025); Zhang et al. (2025) turn such a probe into an early-exit verifier and report about 24% fewer tokens without an accuracy loss. The stopping labels used here are forced-answer correctness, as in Jayabahu & Adeleke (2026), not a learned probe, and training moves the halt into the weights, so decoding needs no auxiliary model.

Domain scope. A descriptive screen of seven non-mathematical domains found median reasoning of 270 to 2,214 tokens (reading-comprehension sets, ARC-Challenge (Clark et al., 2018), MMLU-Pro non-STEM (Wang et al., 2024) and two BBH subtasks (Suzgun et al., 2023)). Mathematics also spans this range: the base model's median is 1,893 think tokens on MATH500 and at least 12,533 on AIME25, and the 24 problems selected for compressibility for the published fit have a median of 5,664. Fitting the recipe on a non-mathematical domain was not attempted: BBH answer extraction and training-set eligibility remained unresolved. The observation was not pre-registered, no verdict rests on it, and all three held-out studies use mathematical problems only.

M FIXED-RECIPE REFIT STABILITY OF THE 22.5 PAIR (OUTCOME-INFORMED, DESCRIPTIVE)

Origin and status. This appendix reports a study that is separate from Studies 1 to 3 and was designed after their results were known. Its pre-registration was committed before any of its outputs existed and states its origin as follows:

> "This fixed-recipe refit study was designed on 2026-09-25, after the Study 3 [...] held-out results, the grader-disagreement cross-tab and the exploratory grader audit [...] had been seen. Its two recipes are the 22.5 pp pair Study 3 selected (R_a21 full-vector reconstruction, S_a11 on-axis reconstruction), and its focus on the 22.5 target, the grader split and missing boxed answers follows directly from those results. The two new training seeds were taken from a roster fixed before any Study 3 outcome existed [...]. Seeds, evaluation draw, sample, endpoints, analysis, technical-failure policy, completion rule, capacity decision rule and stop time were committed before any new output was generated. The study refits fixed recipes and does not repeat the development search or selection, so it measures the stability of these two recipes across refits on the same held-out panel and one evaluation draw, not the reproducibility of the selection procedure. With two new seeds plus one same-seed refit per recipe it gives a descriptive range, not a seed-population estimate."

The quotation describes the pre-registered design; under the capacity rule described below, only the first new seed was scheduled. The pre-registration quotes this passage verbatim; here three internal record identifiers in it are elided and marked [...]. The study refits two fixed recipes. It does not repeat Study 3's development search or selection, it is not a test of Study 3's selection procedure, and it changes no estimator, grader, state, cell or verdict of Study 3.

Design. The two recipes are Study 3's selected 22.5 pair: full-vector reconstruction at $\alpha = 21$ (R) and on-axis reconstruction at $\alpha = 11$ (S), with data, teacher, direction, model, every hyperparameter and the training route unchanged. The pre-registration fixed two new training seeds (3049121431 and 1390311902, the first entries of a seed roster fixed before any Study 3 outcome existed), a same-seed refit control at the original seed 20260918, and a capacity rule, also pinned before launch, that chose from measured GPU capacity and the clock alone between both new seeds, the first new seed only, or no launch. At the decision (2026-09-25 22:08:47 UTC), the sustained capacity was 11 four-GPU evaluators, and the planned first submission time 2026-09-25 22:30:00 UTC was later than the latest time that allowed both new seeds (2026-09-25 21:18:00 UTC) and no later than the latest time that allowed one (2026-09-25 23:31:00 UTC). The rule therefore chose one new seed (3049121431); seed 1390311902 was not scheduled and has no result. The rule was applied twice before launch, both times choosing one new seed, and the later record governs; the choice was recorded before any output of the study existed and was not revisited afterwards. The scheduled work was 4 fits and 83 evaluation jobs: for each scheduled unit and each recipe, one fit and 20 evaluation jobs of 100

problems (40 jobs per unit), plus 3 reproducibility-control jobs. Each refit was evaluated once on all 2,000 held-out problems of Study 3 with the evaluation draw of Study 3's held-out draw seed 173 (four outputs per problem), the same prompt, decoder, caps and graders; only the adapter differs. The pinned comparator is the original checkpoints' values on that same draw (the draw-173 rows of Appendix F), not the four-draw primary result, which has four times as many outputs per problem. Every scheduled unit is reported below, whatever it shows; nothing is pooled across seeds.

What the training seed changes. On the training route used, the training seed sets the LoRA initialization, the data order and dropout jointly, so no statement here isolates any one of them. Fit receipts, read before any evaluation job was submitted, showed that the refit at the original seed reproduced the original initialization exactly. Its final adapter file differed in bytes from the original's; under the pinned rule this is read as training nondeterminism at a fixed seed (a byte-level comparison, not decomposed further). Following the pinned wording rule, every difference between a new seed and the original is described as "refit variation (seed plus training nondeterminism)", and the same-seed refit's difference from the original is shown beside it as its scale. Because re-evaluating an unchanged original checkpoint on the same draw did not reproduce its outputs byte for byte (0 of 400 on the control block), each difference below, including the same-seed refit's, also contains evaluation-path variation; no difference isolates the training seed.

Reproducibility control and reuse of the base-model outputs. Before any refit output was opened, a pre-registered control re-evaluated the first of the fixed blocks (100 problems, chosen by position) at the same draw for the base model and the two original checkpoints, and compared each rerun with the original outputs by a paired interval on four per-problem quantities (thinking length, Math-Verify accuracy, legacy accuracy and closure; 12 cells, familywise level 5%). The base-model rerun was byte-identical to the original in 400 of 400 outputs. The rerun of the original full-vector checkpoint was byte-identical to the original in 0 of 400 outputs. The rerun of the original on-axis checkpoint was byte-identical to the original in 0 of 400 outputs. Byte-level differences were expected on this evaluation path and do not by themselves invalidate anything; the pinned distributional rule decides reuse. No cell showed drift, and the pinned verdict was `REUSE-VALID`: the original base-model outputs are reused as the base for every unit. With 100 problems per job the control detects only large drift (on this block, two independent original draws give intervals spanning about -175 to $+732$ thinking tokens for the base model and about ± 5 to 8 pp on the rates), so reuse means that no gross drift was detected, not that the rerun matches.

Endpoints and states. All estimators, bounds, margins and state rules are Study 3's frozen ones applied to one draw (Appendix B): C_R and C_S are the percentage reductions in summed per-problem median thinking length against the base model, $D = C_R - C_S$, and accuracy and closure contrasts are in pp. Intervals are the frozen paired problem bootstrap with multiplicity divisor 3; one draw has a quarter of the outputs per problem of the four-draw analysis, so its intervals are wider. Matched shortening is retained at a unit if and only if its matching state on D is P and both cut states are P (the frozen rule); otherwise it is not retained, and the failing predicates are named. On this single draw the original checkpoints also read U. The one-draw interval for D that the matching rule uses (adjusted one-sided; Table 35 prints the wider two-sided bounds) is 2.53, 2.50 and 2.62 pp wide at the original checkpoints, the same-seed refit and the new seed, respectively, which is wider than the 2-pp matching band, so no unit could read P on one draw. 'Not retained (match U)' reflects this one-draw resolution limit; it is not evidence that matching was lost, and it is not evidence that matching held. Matching at 22.5 was established by the four-draw analysis (Section 6.3), which is context only and is not a comparator here. The point $|D| \leq 1$ pp is reported separately and is descriptive. Values are printed to four decimals, as in the analysis record. Table 35 gives every unit's values beside the comparator and Table 36 the differences from the comparator. No further adjustment is made for the number of units or for the single draw; the states are descriptive and are not new tests.

Table 35: Refit stability of the 22.5 pair, draw seed 173 only: points with adjusted two-sided bounds in brackets (pp), states, and counts per arm (of 8,000 outputs). The original checkpoints' column is the comparator. On one draw the matching state is U for every column, including the comparator (see text). Accuracy states: E within the 2.6-pp margin, R or S resolved toward full-vector or on-axis reconstruction, U unresolved. "Unboxed MV-only, closed" counts outputs that close the thinking block, contain no boxed answer, and are credited by Math-Verify but not by the legacy grader. Descriptive; no test across seeds.

Endpoint	Original (comparator)	Refit, seed 20260918	Seed 3049121431
C_R	23.6040 [22.0180, 25.1469]	23.6507 [22.0438, 25.2278]	23.5035 [21.9549, 25.0299]
C_S	24.2156 [22.5985, 25.7811]	23.9222 [22.3456, 25.4443]	23.0270 [21.3483, 24.6427]
$D = C_R - C_S$	−0.6116 [−2.0336, 0.8112]	−0.2715 [−1.6776, 1.1363]	0.4765 [−0.9823, 1.9619]
MV, R−base	0.1125 [−0.9250, 1.1625]	−0.2125 [−1.2250, 0.8000]	0.6000 [−0.4125, 1.6125]
MV, S−base	0.1000 [−0.9375, 1.1250]	0.2250 [−0.8625, 1.3000]	0.3625 [−0.6750, 1.3875]
MV, R−S	0.0125 [−0.9750, 1.0000]	−0.4375 [−1.4000, 0.5125]	0.2375 [−0.6500, 1.1375]
legacy, R−base	0.5500 [−0.4875, 1.6000]	0.3000 [−0.7000, 1.3000]	0.7875 [−0.2250, 1.8000]
legacy, S−base	−2.7125 [−3.8750, −1.5625]	−2.0000 [−3.1625, −0.8500]	−2.3500 [−3.4875, −1.2250]
legacy, R−S	3.2625 [2.1750, 4.3625]	2.3000 [1.2875, 3.3250]	3.1375 [2.1250, 4.1625]
closure, R−base	7.0125 [6.0125, 8.0500]	6.7250 [5.7375, 7.7375]	6.6250 [5.6500, 7.6375]
closure, S−base	7.5125 [6.4875, 8.5750]	7.5625 [6.5375, 8.6250]	7.4625 [6.4250, 8.5250]
closure, R−S	−0.5000 [−1.0375, 0.0250]	−0.8375 [−1.3750, −0.3125]	−0.8375 [−1.3625, −0.3250]
Matching state on D	U	U	U
Cut states R/S	P/P	P/P	P/P
Accuracy state MV/legacy	E/R	E/R	E/R
Quality R/S	P/U	U/U	U/U
Flags	GRADER-SPLIT	GRADER-SPLIT	GRADER-SPLIT
Cell	MATCH-UNRESOLVED	MATCH-UNRESOLVED	MATCH-UNRESOLVED
Matched shortening retained	no (match U)	no (match U)	no (match U)
$\lvert D\rvert \le 1$ pp (point)	yes	yes	yes
MV-only R/S/base	929/1190/964	923/1143/964	950/1181/964
Legacy-only R/S/base	0/1/0	0/1/0	1/0/0
No boxed answer R/S/base	261/640/764	303/613/764	290/631/764
No boxed answer, closed R/S/base	82/495/26	101/475/26	91/480/26
Unboxed MV-only, closed R/S/base	48/296/16	55/268/16	56/300/16

Table 36: Differences from the comparator (the original checkpoints, draw seed 173), points only (pp). The same-seed refit's difference is shown first, as the scale against which each new seed's difference is read. The last column is the range of the values themselves (not of the differences) over the comparator and every analysed unit. Descriptive.

Endpoint	Refit (scale) − original	Seed 3049121431 − original	Range of values
C_R	0.0467	−0.1005	[23.5035, 23.6507]
C_S	−0.2934	−1.1886	[23.0270, 24.2156]
$D = C_R - C_S$	0.3402	1.0882	[−0.6116, 0.4765]
MV, R−base	−0.3250	0.4875	[−0.2125, 0.6000]
MV, S−base	0.1250	0.2625	[0.1000, 0.3625]
MV, R−S	−0.4500	0.2250	[−0.4375, 0.2375]
legacy, R−base	−0.2500	0.2375	[0.3000, 0.7875]
legacy, S−base	0.7125	0.3625	[−2.7125, −2.0000]
legacy, R−S	−0.9625	−0.1250	[2.3000, 3.2625]
closure, R−base	−0.2875	−0.3875	[6.6250, 7.0125]
closure, S−base	0.0500	−0.0500	[7.4625, 7.5625]
closure, R−S	−0.3375	−0.3375	[−0.8375, −0.5000]

Per unit. The same-seed refit control (seed 20260918): matched shortening was not retained under the frozen rule: the matching state on D was U (unresolved). As at the comparator, this reflects the one-draw resolution limit described under Endpoints and states; it is evidence neither that matching was lost nor that it held. The matching state was not F: the interval did not lie wholly outside the 1-pp band. The point $|D|$ was 0.2715 pp, at most 1 pp (descriptive). The accuracy states were E under Math-Verify and R under the legacy grader (comparator: E and R). By the pinned label, the 22.5 grader split recurs at this unit (its flags include GRADER-SPLIT and both accuracy states equal the comparator's). By the pinned label, the on-axis unboxed MV-only excess recurs: 268 on-axis against 55 full-vector outputs that close the thinking block, contain no boxed answer and are

credited by Math-Verify alone (comparator, on-axis minus full-vector: 248). States that differ from the comparator's: `closure_RS`, `Q_R`.

The new seed 3049121431: matched shortening was not retained under the frozen rule: the matching state on D was U (unresolved). As at the comparator, this reflects the one-draw resolution limit described under Endpoints and states; it is evidence neither that matching was lost nor that it held. The matching state was not F: the interval did not lie wholly outside the 1-pp band. The point $|D|$ was 0.4765 pp, at most 1 pp (descriptive). The accuracy states were E under Math-Verify and R under the legacy grader (comparator: E and R). By the pinned label, the 22.5 grader split recurs at this unit (its flags include GRADER-SPLIT and both accuracy states equal the comparator's). By the pinned label, the on-axis unboxed MV-only excess recurs: 300 on-axis against 56 full-vector outputs that close the thinking block, contain no boxed answer and are credited by Math-Verify alone (comparator, on-axis minus full-vector: 248). States that differ from the comparator's: `closure_RS`, `Q_R`.

Across units. Over the comparator and the analysed units (original checkpoints, same-seed refit, seed 3049121431), the points take both signs for: `D`, `mv_RB`, `mv_RS`, `mv_SR`. A state is unresolved at some of them and resolved at others for: `closure_RS`, `Q_R`. These two labels are descriptive; no interval, mean, spread or test across seeds is computed.

What this study does and does not show. At most, it shows the per-unit values and states above for the 22.5 pair, on one evaluation draw of the frozen held-out problems, beside the original checkpoints' values on the same draw; whether each unit retained matched shortening under the frozen rule; and whether the grader split and the on-axis unboxed MV-only excess recur by the pinned labels. Because a single draw cannot resolve matching at these interval widths, the pre-registered retention criterion could not be met by any unit, including the original checkpoints; the study can show whether matching is contradicted at a unit (it is not at either), not whether it is retained. It does not repeat Study 3 or its selection procedure, and it gives no seed-population estimate, interval, mean or significance: with at most one new-seed refit and one same-seed refit per recipe it gives a descriptive range. Agreement at a seed is not evidence for Study 3's result, and disagreement at a seed is not evidence against it; the number of refits is too small for either reading. Nothing here changes Study 3's frozen verdicts, estimators, graders or primary endpoint, says anything about the 12.5 or 17.5 targets, other problems, models, directions or prompts, or prefers either grader.

N Development formatting across all recipes (post hoc, exploratory)

Question and status. After the Study 3 readout and the grader audit, we asked whether the on-axis excess of closed, unboxed answers is specific to the selected checkpoints. We used Study 3's development outputs for all 46 recipes and the base model, 3,200 outputs each (200 development problems, four draws of four samples). The analysis is post hoc, exploratory and descriptive. It uses the stored grader verdicts and the literal box rule of Section 6.4, reruns no grader, adjudicates no answer, and changes no feasibility decision, selection, state or cell. Its plan was saved before any new cross-tabulation, and its counts were reproduced by a second count written by the same analyst and by a third, independent implementation.

Definitions. Omission is the absence of the literal string `\boxed` anywhere in an output, and omission among closed outputs divides by the outputs that close `</think>`. Closed, unboxed, MV-only outputs close the thinking block, contain no box, and are credited by Math-Verify but not by the legacy grader; this is stored grader credit, not an assessment of the stated answer. Grader disagreement counts outputs on which the two graders differ.

Findings. Table 37 lists every recipe. At all 23 same-α pairs the on-axis recipe has a higher omission rate among closed outputs (by 0.19 to 6.87 pp), more closed, unboxed, MV-only outputs and more grader disagreement than the full-vector recipe. The differences are small at $\alpha = 5$ to 7 and larger from $\alpha = 8$ on, including the feasible on-axis recipes at $\alpha = 8$ to 11, whose omission among closed outputs is 3.37 to 5.55% against 0.20 to 0.33% for the full-vector recipes at the same α. Total omission, which also counts unclosed outputs, is lower for the on-axis recipes at $\alpha = 5$ to 7. The pattern is therefore not specific to recipes that failed feasibility. All 16 on-axis recipes with $\alpha \geq 12$

failed development feasibility, each on legacy accuracy; pooled over them, omission among closed outputs is 6.41%, against 1.07% for the 16 full-vector recipes at the same α. Two of those on-axis recipes ($\alpha = 35$ and 50) also fail Math-Verify retention, and $\alpha = 50$ also fails closure; these data do not establish that the feasibility failures are purely formatting-related.

Limitations. Legacy accuracy is itself a feasibility criterion, so the association between legacy-failing recipes and higher Math-Verify-only disagreement is partly definitional. Same-α pairs are not matched on shortening (the on-axis minus full-vector development cut ranges from -5.69 to $+8.22$ pp), and α changes the loss normalization as well as the target, so the table is neither a matched comparison nor a causal dose response. Each recipe has one training seed and all recipes reuse the same 200 development problems, so pooled rates are repeated, dependent observations, not independent samples. The $\alpha \geq 12$ grouping was chosen post hoc. No answer was adjudicated, and the conclusions of the held-out audit do not transfer to these outputs. The analysis does not establish that formatting caused any feasibility failure or that unboxed answers are mathematically correct.

Table 37: Study 3 development formatting for every recipe (post hoc, descriptive; 3,200 outputs per recipe on 200 development problems). Cut: development cut in percent of base think tokens. Feasible: the frozen development feasibility state. Omission among closed: outputs without `\boxed` among those that close `</think>`. Closed, unboxed, MV-only and grader disagreement are counts. Same-α rows are not matched on shortening.

	Cut (%)		Feasible		Omission among closed (%)		Closed, unboxed, MV-only		Grader disagreement	
α	R	S	R	S	R	S	R	S	R	S
5	6.25	10.43	yes	yes	0.30	0.49	6	8	373	381
6	9.32	13.16	yes	yes	0.33	0.62	3	13	378	395
7	10.25	15.27	yes	yes	0.33	1.56	7	30	377	397
8	10.37	17.86	yes	yes	0.20	3.37	3	69	381	441
9	11.91	19.60	yes	yes	0.33	5.11	7	105	380	483
10	12.45	20.16	yes	yes	0.20	4.31	4	81	370	441
11	14.47	22.69	yes	yes	0.26	5.55	5	110	372	488
12	15.22	22.68	yes	no	0.59	5.46	11	96	388	463
13	15.29	22.03	yes	no	0.39	6.06	6	116	381	492
14	17.47	23.29	yes	no	0.49	4.74	7	102	374	478
15	18.01	24.14	yes	no	0.52	5.72	13	131	391	488
16	18.51	24.08	yes	no	0.55	6.22	12	129	382	491
17	19.99	25.63	yes	no	0.32	6.53	7	140	377	503
18	20.13	25.86	yes	no	0.45	6.54	8	134	369	499
19	20.99	26.46	yes	no	0.61	6.82	10	146	386	502
20	21.93	25.80	yes	no	1.00	6.77	18	142	382	504
21	22.49	27.23	yes	no	0.84	7.69	20	150	385	515
22	22.84	25.67	yes	no	0.83	7.71	20	166	388	527
23	23.48	26.60	yes	no	1.32	7.31	23	155	390	533
24	24.99	26.41	yes	no	1.06	7.10	20	147	370	514
25	25.33	27.82	yes	no	1.06	6.83	22	133	389	489
35	31.25	29.20	no	no	3.55	5.71	76	122	444	483
50	31.56	25.87	no	no	3.38	5.17	68	92	430	460
Base	reference				0.28		1		378	