WHAT DOES PRIVILEGED INFORMATION ADD TO ON-POLICY SELF-DISTILLATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) lets a language model learn from a frozen copy of itself that sees an answer or a worked solution. Giving the teacher this extra information seems to offer the student more to learn, but how much does it add beyond distillation itself? To isolate that contribution, we construct AMPLE-MATH, a reusable suite of 5,319 mathematical problems with six reasoning views that share the same answer, and compare each view with matched reference-free distillation. With a thinking-enabled teacher supervising direct-response rollouts, reference-free distillation accounts for much of Qwen3-1.7B's improvement under thinking-enabled evaluation, both in domain and on external benchmarks. Evidence for an additional reference benefit is modest in Qwen, strongest for a polished solution, whereas complete traces add two percentage points in SmolLM3-3B at step 50. These benefits depend on the student being trained. At the same checkpoint, replacing short direct-response rollouts with long thinking-enabled rollouts turns gains into losses in both families while the problems, references, and evaluation stay fixed. Teacher profiles and matched loss interventions in Qwen further show that changing token-level supervision can leave student behavior largely unchanged. Together, these findings suggest that OPSD can improve access to existing reasoning capabilities through parameters shared by direct-response and thinking-enabled inference. The value of a privileged reference is what it adds to this cross-mode transfer, not how much of the solution it reveals.

1 INTRODUCTION

A worked solution gives a teacher information that its student does not have. On-policy self-distillation (OPSD) uses this asymmetry to improve a language model without a separate, larger teacher. A frozen copy of the model sees the solution and scores responses generated by the student, which sees only the problem (Agarwal et al., 2024; Zhao et al., 2026a). As in learning with privileged information(PI) (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016), the extra information supports training without being needed at inference. Similarly, the best content and form of privileged information for this type of training remains undecided. A fuller reference reveals more about the solution, but does it give the student more to learn?

Improvement over the base model alone does not answer that question. Recent studies find useful supervision with absent or other-problem references (Shrestha & Tessier, 2026; Ichihara et al., 2026), while structured guidance can outperform complete solutions (Zhao et al., 2026b). These findings establish the need to separate what a reference contributes from what distillation already achieves. We build on these controls by holding the answer fixed across reasoning representations and tracking their effects as the student changes. In OPSD, the teacher scores the student's own prefixes rather than supplying a solution to reproduce. A reference must therefore help the teacher give useful feedback on the student's attempt.

To study this interaction while holding answer information fixed, we construct AMPLE-MATH (Answer-Matched Privileged Levels of Explanation) from OpenThoughts-114k (Guha et al., 2026). Each of its 5,319 problems has six views, from ANSWER ONLY to FULL TRACE, sharing one verified answer while varying the reasoning representation. We compare each view with a reference-free control under the same training configuration (Figure 1). A thinking-enabled teacher supervises direct-response rollouts (Zhao et al., 2026a; Ichihara et al., 2026), and we evaluate the student with

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

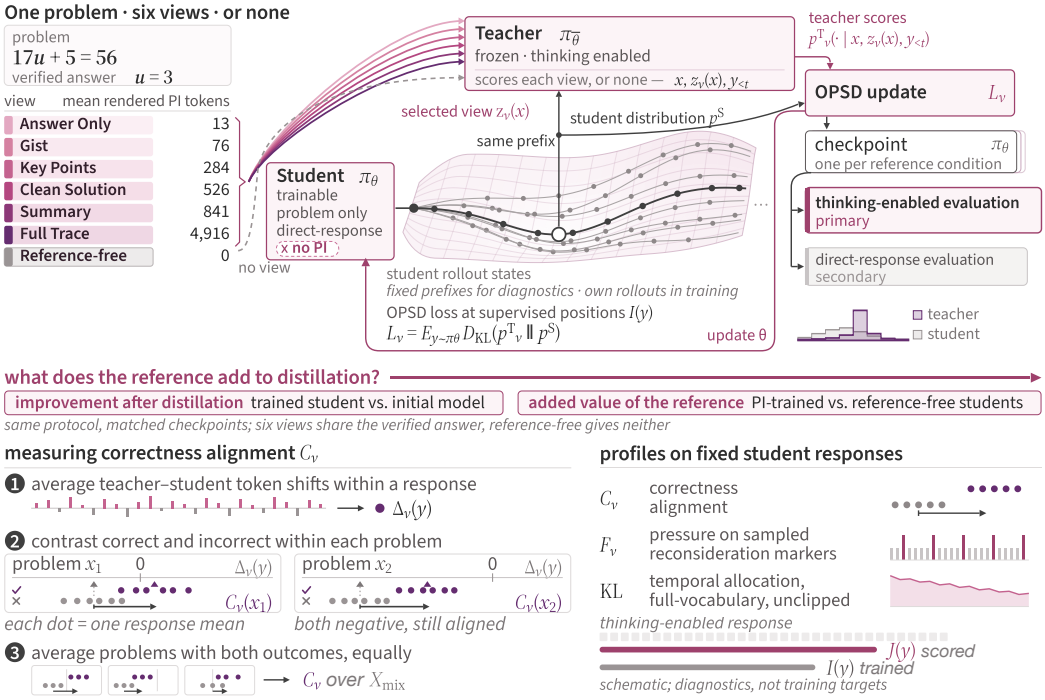


Figure 1: **What does the reference add to distillation?** The frozen, thinking-enabled teacher scores student prefixes with or without privileged information. Lower panels illustrate correctness alignment and diagnostic profiles. The loss is schematic, with clipping specified in Equation 1.

thinking enabled. Comparisons in Qwen3-1.7B (Yang et al., 2025) and SmolLM3-3B (Bakouch et al., 2025) follow the reference's contribution across checkpoints.

Removing the reference does not fully remove the teacher–student asymmetry. The original OPSD configuration and several subsequent methods pair a thinking-enabled teacher with a direct-response student (Zhao et al., 2026a; Lin et al., 2026; Kara & Ersoy, 2026; Li et al., 2026b). As a result, the thinking-enabled teacher still scores prefixes produced with thinking disabled even without any PI (Ichihara et al., 2026; Shrestha & Tessier, 2026). Learning from these scores changes parameters shared by both student inference modes, offering a route to stronger thinking-enabled reasoning without additional solution information. Much of Qwen's improvement is already present without a reference, with modest evidence of an additional benefit from CLEAN SOLUTION. In SmolLM3, FULL TRACE adds two points beyond reference-free training at step 50. The same references nevertheless accompany gains under direct-response training rollouts and losses under long thinking-enabled rollouts in both families. What the reference contributes therefore depends on the student responses it helps the teacher supervise.

To understand this dependence, we examine what the teacher's scores encourage on student responses. Prior work points to preferences for the supplied solution (Harne et al., 2026), likelihood shifts on student tokens (Nguyen et al., 2026), and suppressed reconsideration (Kaur et al., 2026) as possible sources of poor transfer. Our Qwen profiles show that correctness alignment changes with the responses being scored, while sampled correction markers receive negative pressure in both prefix modes. Relaxing loss at these markers barely changes their use, and an apparent benefit from broader supervision mostly disappears when checkpoints are matched. Teacher profiles identify explanations to test, while matched training outcomes determine which changes help.

Our contributions can be summarized as follows. First, we construct AMPLE-MATH, a reusable supervision suite for isolating reasoning representation through six answer-matched views per problem, with audits and frozen splits. Second, controlled comparisons in two model families separate the benefits of cross-mode distillation from a reference's added value and show that changing student training trajectories can reverse transfer. Third, teacher profiles and matched interventions

distinguish changes to supervision from changes in student behavior, and show that stopping time explains most of an apparent gain from broader loss coverage.

2 A Controlled Study of Privileged Supervision

Our study separates the contribution of a privileged reference from the gains of distillation, using thinking-enabled student performance as the primary outcome.

2.1 OPSD on Student-Generated Prefixes

For a problem x, view v supplies PI $z_v(x)$ to a frozen teacher $\pi_{\bar{\theta}}$. The student π_θ starts from the same backbone and generates $y = (y_1, \ldots, y_L)$ from the problem alone under rollout configuration m_S. For a vocabulary token a at prefix $s_t = (x, y_{<t})$, the student's next-token distribution is $p_t^S(a) = \pi_\theta(a \mid x, y_{<t}; m_S)$. The teacher scores the same prefix using $p_{v,t}^T(a) = \pi_{\bar{\theta}}(a \mid x, z_v(x), y_{<t}; m_T, \tau)$. Here m_T specifies the teacher's reasoning configuration and τ its scoring temperature. The reference-free control omits the PI section but retains the thinking-enabled teacher scoring direct-response student prefixes. It is therefore cross-mode self-distillation, not a comparison of identical teacher and student distributions. For the supervised positions $I(y)$, the OPSD loss is

$$\mathcal{L}_v(\theta; m_S) = \mathbb{E}_{x,\, y \sim \pi_\theta(\cdot|x; m_S)} \left[\frac{1}{|I(y)|} \sum_{t \in I(y)} D_{\mathrm{gKL}}\big(p_{v,t}^T \,\|\, p_t^S\big) \right], \tag{1}$$

where D_{gKL} clips each vocabulary term in the teacher-to-student forward KL from above at 0.05 before summation (Zhao et al., 2026a). The teacher distribution and sampled completion are held fixed when differentiating the loss. With the same prompt, thinking mode, initial parameters, and scoring temperature, a reference-free teacher would match the student and give zero initial loss. PI then supplies the initial discrepancy, while later updates also separate the student's parameters from the frozen teacher.

2.2 What a Privilege Profile Measures

Before training, we profile Qwen3-1.7B's token-level supervision on 512 problems, using four unprivileged completions per problem under each rollout configuration. For each completion y, let $J(y)$ index the longest prefix fitting every teacher context, capped at 1,024 completion tokens for direct-response profiles. Scoring the same tokens under every view gives the mean log-probability shift

$$\Delta_v(y) = \frac{1}{|J(y)|} \sum_{t \in J(y)} \big[\log p_{v,t}^T(y_t) - \log p_t^S(y_t) \big]. \tag{2}$$

Let $\mathcal{Y}_x^+$ and $\mathcal{Y}_x^-$ contain the correct and incorrect profiled samples for problem x, and let $\mathcal{X}_{\mathrm{mix}}$ contain problems with both outcomes. **Correctness alignment** C_v contrasts responses for each $x \in \mathcal{X}_{\mathrm{mix}}$, then averages these problems equally:

$$C_v(x) = \frac{\sum_{y \in \mathcal{Y}_x^+} \Delta_v(y)}{|\mathcal{Y}_x^+|} - \frac{\sum_{y \in \mathcal{Y}_x^-} \Delta_v(y)}{|\mathcal{Y}_x^-|}, \qquad C_v = \frac{1}{|\mathcal{X}_{\mathrm{mix}}|} \sum_{x \in \mathcal{X}_{\mathrm{mix}}} C_v(x). \tag{3}$$

Positive alignment means larger teacher–student shifts for correct than incorrect responses, on average within problems. **Correction pressure** F_v averages the shift at sampled reconsideration markers such as *wait*, *but*, and *check*, with negative values indicating lower marker probability. **Temporal KL allocation** is the share of full-vocabulary $D_{\mathrm{KL}}(p_{v,t}^T \,\|\, p_t^S)$ in each quarter of the retained span. We use unclipped KL and normalize quarters within this span, not the training loss support.

Holding student responses fixed lets us compare how different references shape supervision on the same prefixes. Correctness alignment captures whether correct responses receive more favorable shifts than incorrect ones, while marker pressure and KL allocation guide our tests of marker weighting and loss coverage. We then measure transfer after training and repeat the profiles on trained students' responses to examine which supervision patterns persist as the student learns. Appendix D.4 derives the local reweighting interpretation of correctness alignment.

2.3 AMPLE-MATH: ANSWER-MATCHED PRIVILEGED VIEWS

We construct AMPLE-MATH by pairing 5,319 mathematical problems from OpenThoughts-114k (Guha et al., 2026) with six answer-matched views. ANSWER ONLY supplies no reasoning body, GIST the central method, and KEY POINTS ordered steps. CLEAN SOLUTION supplies a polished solution, SUMMARY a narrative summary, and FULL TRACE the complete source trace. Every view ends with the same canonical answer section. Mean rendered lengths range from 13 to 4,916 Qwen3 tokens, describing reasoning density rather than assumed quality. GIST, KEY POINTS, and SUMMARY are generated from the intact source trace by Qwen3.6-35B-A3B-FP8 (Qwen Team, 2026) and checked for source fidelity by the same model under a separate review prompt. ANSWER ONLY is deterministic, while CLEAN SOLUTION and FULL TRACE are source-derived.

We annotate problem difficulty using 8–16 problem-only samples per question from each of three evaluator models and a binomial Rasch model (Rasch, 1960). Separately, four unprivileged direct-response samples per problem from the frozen Qwen3-1.7B base define three model-relative success bands for the experimental splits: three or four correct, one or two correct, and none correct. The zero-success band requires a correct generation from a structured-PI teacher view. Equal band quotas give train, development, and test sets of 1,536, 192, and 384 problems. SmolLM3 uses the same development and test problems but a training split rebuilt from its own direct-response outcomes. Appendix B details construction, filtering, fidelity audits, difficulty estimation, and splits.

2.4 TRAINING AND EVALUATION

Models and rollouts. The main experiments train Qwen3-1.7B with LoRA adapters (Hu et al., 2022) ($r = 64$, $\alpha = 128$) for 100 optimizer steps, with SmolLM3-3B providing the cross-family test (Yang et al., 2025; Bakouch et al., 2025). The teacher is always the frozen, thinking-enabled backbone. Our primary configuration uses direct-response training followed by thinking-enabled evaluation, an asymmetry also examined in other OPSD studies (Zhao et al., 2026a; Ichihara et al., 2026; Shrestha & Tessier, 2026). The student generates at most 1,024 tokens with thinking disabled, while the thinking-enabled training comparison extends rollout generation.

Evaluation and statistics. Primary evaluation uses four unprivileged thinking-enabled samples per problem. Secondary comparisons evaluate the same checkpoints with thinking disabled. Avg@4 averages the fraction of correct samples across problems. In-domain evaluation uses a 16,384-token budget and regenerates capped responses with the same sampling seed and up to 32,512 tokens. The length analyses grade the first 4,096, 8,192, and 16,384 saved token IDs from the initial pass, rather than generating new responses at those limits. External evaluation uses 12 thinking-enabled samples per problem (Zhao et al., 2026a) on AIME 2024 and 2025 (MAA, 2024; 2025) and HMMT February 2025 (HMMT, 2025).

The principal reference comparisons use steps 50 and 100 in both families. Paired 95% intervals use 10,000 problem-cluster bootstrap resamples. For the initial six-view control tests, we compare seed-0 students with three-seed controls and apply Holm correction across the six contrasts. Follow-up seed averages are labeled in the figures and tables. Their intervals resample problems after averaging the observed seeds, with seed-level t intervals reported as a separate check. Full profiling, training, and evaluation details appear in Appendix A.

3 WHAT PRIVILEGED REFERENCES ADD

Improvement over the base model can come from distillation itself, so we compare students trained with and without a teacher reference to identify what privileged information adds.

3.1 DISTILLATION GAINS AND REFERENCE CONTRIBUTIONS

In Qwen, much of the improvement attributed to privileged supervision is already present without a reference (Figure 2a, Table 1). The reference-free student improves both in domain and on external benchmarks, and even the wrong-answer control improves over the base. More reasoning brings no consistent increase in gains. At step 100, ANSWER ONLY and FULL TRACE sit about 0.6 points above the reference-free student, with both intervals including zero.

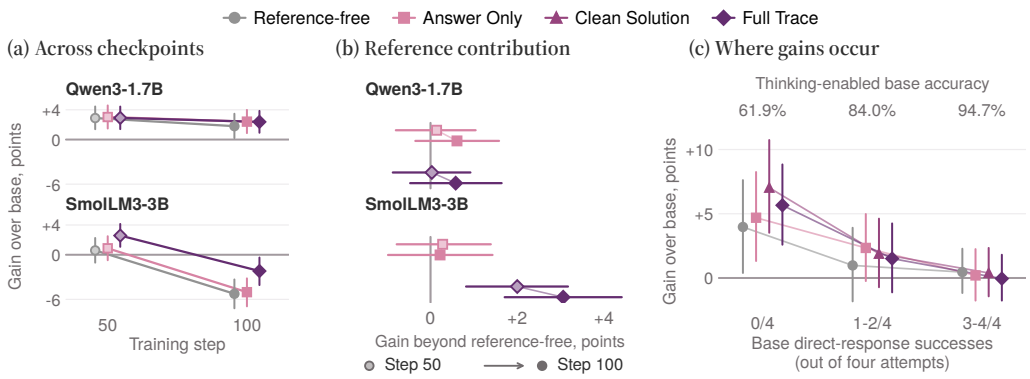


Figure 2: **Distillation gains and reference contributions are different quantities.** Direct-response training and thinking-enabled evaluation throughout. (a) Gains over the base across checkpoints. (b) Paired gains beyond reference-free training. (c) Qwen step-100 gains grouped by base direct-response successes. Top annotations give thinking-enabled base accuracy. Error bars show unadjusted 95% problem-cluster bootstrap confidence intervals (384 problems, 128 per subgroup). Three seeds per configuration, except four for Qwen ANSWER ONLY/FULL TRACE.

Table 1: **Six-view transfer in domain and on external benchmarks.** Qwen3-1.7B gains over each protocol's frozen base after direct-response training, averaged over the listed training seeds (in/ex denotes in-domain/external). Wrong answer uses ANSWER ONLY with another problem's answer. All evaluation is thinking-enabled, with 30 problems per external benchmark. Subscripts give 95% problem-cluster bootstrap intervals. Green/red shading shows positive/negative gains on a shared ± 10-point scale.

			In domain	**External, Δ Avg@12 at step 50**			
Reference	**PI tokens**	**Seeds (in/ex)**	**Step 100 Δ Avg@4**	**Pooled**	**AIME 2024**	**AIME 2025**	**HMMT 2025**
Reference-free	0	3/3	$+1.80_{[0.20,3.47]}$	$+4.07_{[1.73,6.54]}$	$+8.24_{[3.61,13.15]}$	$+1.67_{[-1.85,5.37]}$	$+2.31_{[-1.39,6.11]}$
Wrong answer	13	3/3	$+1.67_{[0.07,3.32]}$	$+4.14_{[1.33,7.07]}$	$+6.30_{[0.28,12.59]}$	$+2.87_{[-1.11,6.94]}$	$+3.24_{[-1.02,7.50]}$
ANSWER ONLY	13	4/4	$+2.41_{[0.86,3.99]}$	$+3.73_{[1.27,6.41]}$	$+8.06_{[3.61,13.06]}$	$+1.53_{[-1.18,4.44]}$	$+1.60_{[-3.06,7.01]}$
GIST	76	1/1	$+2.15_{[0.26,4.04]}$	$+4.35_{[1.30,7.59]}$	$+6.67_{[0.56,13.06]}$	$+3.89_{[-0.56,8.33]}$	$+2.50_{[-3.06,8.06]}$
KEY POINTS	284	1/1	$+2.80_{[0.72,4.88]}$	$+3.61_{[0.37,7.04]}$	$+3.89_{[-2.78,10.56]}$	$+5.00_{[0.28,9.72]}$	$+1.94_{[-3.33,7.78]}$
CLEAN SOLUTION	526	3/1	$+3.10_{[1.54,4.73]}$	$+4.91_{[2.04,7.96]}$	$+7.78_{[2.22,13.89]}$	$+4.44_{[0.83,8.89]}$	$+2.50_{[-2.50,8.06]}$
SUMMARY	841	1/1	$+2.54_{[0.71,4.43]}$	$+3.61_{[0.83,6.48]}$	$+6.94_{[1.39,12.78]}$	$+1.67_{[-2.50,6.11]}$	$+2.22_{[-1.94,6.67]}$
FULL TRACE	4,916	4/4	$+2.38_{[0.91,3.86]}$	$+2.29_{[-0.16,4.88]}$	$+5.97_{[1.39,10.97]}$	$+0.49_{[-3.47,4.65]}$	$+0.42_{[-3.12,4.24]}$

CLEAN SOLUTION gives modest evidence that a reference can add value in Qwen. Across three seeds per configuration, its step-100 advantage over reference-free training is $1.30\,[0.20, 2.41]$ points. The interval excludes zero before adjustment, but the effect does not survive Holm correction across six views. Thus the evidence favors a small, view-specific contribution rather than a general benefit from supplying more of the solution.

Gains are largest on problems the base never solved in four direct-response attempts, but solves 62% of the time with thinking enabled (Figure 2c). This pattern holds with and without PI, and CLEAN SOLUTION's additional benefit is also largest in this group. In the separate original-data validation, external gains concentrate on problems solved in some but not all of the base's twelve thinking-enabled attempts. Both patterns fit the shared-parameter interpretation. Learning from thinking-enabled teacher scores on direct-response prefixes may also improve the student's thinking-enabled reasoning.

The reference makes a clearer contribution in SmolLM3, where FULL TRACE adds 2.0 points over reference-free training at step 50 with thinking enabled (Figure 2b). By step 100 all three configurations have fallen below the base, but FULL TRACE loses less than the reference-free

and ANSWER ONLY students. The same reference helps the earlier student and softens the later deterioration, two different contributions that a gain over the base alone would not distinguish.

The same students rank references differently when answering directly. At step 100, Qwen's ANSWER ONLY and reference-free students exceed FULL TRACE by 5.32 and 10.81 points under direct-response evaluation. SmolLM3's step-50 FULL TRACE advantage likewise changes from $+2.00$ points with thinking enabled to -10.63 when answering directly. Qwen's direct-response winners also write longer responses, and grading only their first 4K tokens reverses the ordering. The ranking changes accompany differences in response length, linking reference choice to how the student answers as well as how often it is correct.

3.2 STUDENT TRAINING TRAJECTORIES CHANGE TRANSFER

How the student answers also matters during training, when its responses determine the prefixes the teacher supervises. To examine the thinking-model degradation reported by Kaur et al. (2026), we replace direct-response rollouts with thinking-enabled rollouts under ANSWER ONLY, CLEAN SOLUTION, and FULL TRACE supervision while keeping thinking-enabled evaluation fixed. Thinking-enabled rollouts extend beyond the inherited first-1,024-token loss window, so the reasoning mode, the horizon, and the fraction of the response supervised change together.

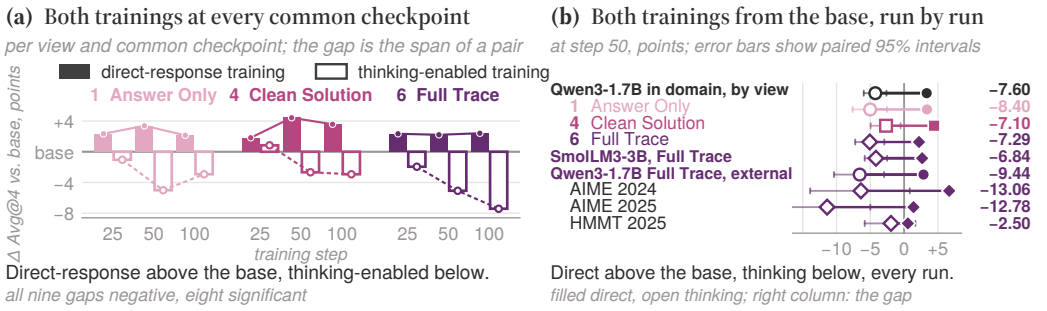


Figure 3: **Changing training trajectories reverses transfer.** Thinking-enabled evaluation throughout. (a) Avg@4 gains over the base across checkpoints. (b) Step-50 gains: Qwen3-1.7B in domain, two-seed SmolLM3-3B under FULL TRACE, and Qwen under FULL TRACE externally. Filled/open marks denote direct-response/thinking-enabled training. Bold rows pool groups. Right-hand contrasts are thinking-enabled minus direct-response, with paired 95% problem-cluster bootstrap intervals.

The same teacher and reference now accompany opposite transfer outcomes. At step 50 the thinking-enabled training configuration turns gains over the base into losses for all three views, with a penalty that persists across checkpoints (Figure 3a). The step-50 FULL TRACE comparison also holds in SmolLM3 and on Qwen's external benchmarks (Figure 3b). The reference is unchanged, but its supervision is now applied to a different kind of student attempt.

4 HOW STUDENT RESPONSES SHAPE SUPERVISION

To interpret the training-trajectory contrast, we profile probability gaps between privileged teachers and unprivileged students. Teacher weights stay fixed, but predictions depend on the reference and the student's prefix.

Correctness alignment changes with the responses being scored. For FULL TRACE, correct responses receive larger average teacher–student log-probability shifts than incorrect ones on direct-response prefixes. The ordering reverses on thinking-enabled prefixes (Figure 4a). On direct-response-trained students' responses, alignment is much smaller for five views, while FULL TRACE retains most of its initial value. These profiles use new responses and updated student distributions. Near-zero alignment means similar average shifts for correct and incorrect responses, not token-level agreement.

Denser references induce stronger overall shifts, which raw C_v does not separate from correctness selectivity. Restricting thinking-enabled profiles to their first 1,024 scored tokens also weakens

differences across views. Alignment thus describes supervision on a particular response distribution and span, not a fixed quality of the reference.

Correction-marker suppression is shared across prefix modes. Kaur et al. (2026) propose suppressed reconsideration as an explanation for thinking-model degradation. In the frozen Qwen base, every view lowers sampled reconsideration-marker probability on average in both prefix modes (Figure 4b). This shared sign does not distinguish the configurations with opposing transfer outcomes, motivating the targeted loss edits below.

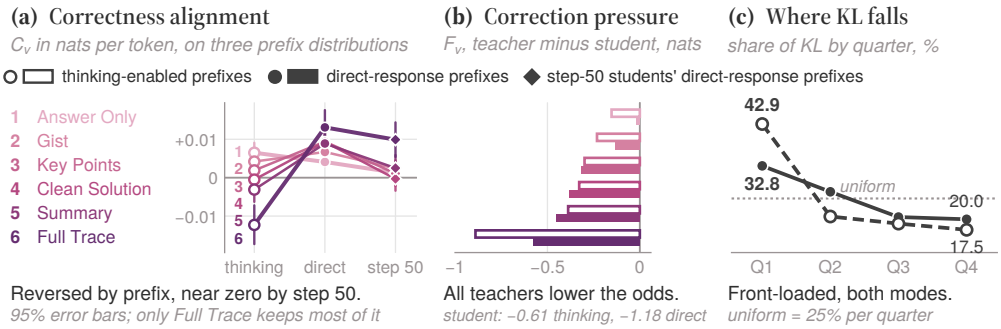


Figure 4: **Teacher profiles separate correctness alignment, marker pressure, and supervision placement.** (a) Correctness alignment C_v, (b) sampled reconsideration-marker pressure F_v, and (c) full-vocabulary KL shares by quarter of the scored span, not the training loss window. Open/filled marks show frozen-base thinking-enabled/direct-response prefixes; diamonds show direct-response-trained students' prefixes at step 50. Error bars show 95% problem-cluster bootstrap intervals.

Diagnostic KL is concentrated near the response opening. The first quarter of the retained span carries disproportionate diagnostic KL in both prefix modes (Figure 4c). This locates teacher–student disagreement, not its association with correctness. The inherited 1,024-token loss window (Zhao et al., 2026a) covers full direct-response rollouts but only the opening of long thinking-enabled rollouts. We therefore test broader loss coverage for the long rollouts.

5 Which Supervision Changes Matter for Transfer

The profiles suggest where supervision might be changed, but only training reveals whether those changes help. We test this in Qwen3-1.7B by modifying reference content and the distillation loss while keeping thinking-enabled evaluation fixed (Table 2).

Reference-free gains do not make reference content irrelevant. Replacing Clean Solution or Key Points with a length-matched view from another problem, including that problem's answer, lowers thinking-enabled accuracy by about two points relative to the genuine view in two single-seed comparisons. Matching the reference's length does not preserve its effect when the text concerns a different problem.

Shortening a reference instead tests how much of a relevant solution is useful. We cut Full Trace to each problem's Key Points, Clean Solution, or Summary length, retaining the trace opening but removing its later reasoning and canonical answer section. These openings bring no clear gain with thinking enabled, although they improve direct-response accuracy over matched Full Trace by roughly 8–11 points. Compression therefore changes how the student benefits from the reference across inference modes.

Relaxing correction-marker loss barely changes marker use. Under direct-response training, we exclude or downweight loss at sampled correction markers for Full Trace and Clean Solution. Neither edit produces an accuracy gain at step 100, and marker use remains almost unchanged. Relaxing a penalty on sampled markers is distinct from encouraging the student to reconsider.

Checkpoint choice explains most of the apparent loss-window gain. Beyond individual marker positions, we test whether long thinking-enabled rollouts benefit from supervision beyond their

Table 2: **Matched interventions on reference content and the training signal.** Seventeen single-seed comparisons, all evaluated with thinking enabled. Upper rows use direct-response training at step 100, lower rows compare thinking-enabled FULL TRACE loss windows at matched steps. Avg@4 is in percent, accuracy changes in percentage points, and subscripts give 95% problem-cluster bootstrap intervals. Δ tokens gives mean length change, and Δ markers the relative change in mean per-response markers per 1,000 tokens, both versus the matched configuration. Trace openings omit the canonical answer section. Bold contrasts pass Holm's multiple-comparison correction within the two other-problem controls. Colors follow Table 1's ±10-point scale.

Intervention	Avg@4	Δ vs. matched	Δ vs. base	Δ tokens	Δ markers
Other-problem reference length-matched view, including the other problem's answer					
CLEAN SOLUTION other-problem view	$81.84_{[78.71,84.83]}$	$\mathbf{-1.95}_{[-3.78,-0.20]}$	$+1.63_{[-0.39,3.65]}$	−794	+4.2%
KEY POINTS other-problem view	$80.86_{[77.67,83.85]}$	$\mathbf{-2.15}_{[-3.97,-0.39]}$	$+0.65_{[-1.24,2.48]}$	−1,242	+7.6%
Trace opening FULL TRACE cut to the length of CLEAN SOLUTION, KEY POINTS, or SUMMARY					
FULL TRACE → CLEAN SOLUTION	$82.03_{[78.97,84.96]}$	$-0.59_{[-2.28,1.11]}$	$+1.82_{[0.13,3.58]}$	+684	+4.3%
FULL TRACE → KEY POINTS	$82.42_{[79.36,85.42]}$	$-0.20_{[-2.02,1.69]}$	$+2.21_{[0.46,4.10]}$	+464	+6.1%
FULL TRACE → SUMMARY	$81.51_{[78.45,84.51]}$	$-1.11_{[-2.86,0.65]}$	$+1.30_{[-0.33,2.93]}$	+599	+2.2%
Prompt templates alternative templates for the teacher and student					
FULL TRACE swap	$80.99_{[77.80,84.05]}$	$-1.63_{[-3.39,0.13]}$	$+0.78_{[-1.17,2.80]}$	+568	−1.5%
CLEAN SOLUTION swap	$83.92_{[80.86,86.85]}$	$+0.13_{[-1.50,1.76]}$	$+3.71_{[2.02,5.47]}$	+1,068	−1.2%
Correction markers loss excluded or downweighted at sampled marker positions					
FULL TRACE exclude	$81.97_{[78.91,84.83]}$	$-0.65_{[-2.41,1.17]}$	$+1.76_{[-0.07,3.58]}$	−164	+0.4%
FULL TRACE downweight	$81.84_{[78.71,84.83]}$	$-0.78_{[-2.34,0.78]}$	$+1.63_{[-0.07,3.32]}$	−64	−0.3%
CLEAN SOLUTION exclude	$83.07_{[79.95,86.00]}$	$-0.72_{[-2.21,0.78]}$	$+2.86_{[1.04,4.69]}$	+231	+0.3%
CLEAN SOLUTION downweight	$83.07_{[80.08,86.00]}$	$-0.72_{[-2.34,0.91]}$	$+2.86_{[1.04,4.69]}$	+132	+1.1%

Loss support thinking-enabled FULL TRACE, each alternative minus EARLY-1K	**Step 25**	**Step 50**	**Step 100**
First-4K	$+0.98_{[-0.91,2.86]}$	$-0.85_{[-2.93,1.17]}$	$-1.56_{[-3.91,0.72]}$
Distributed-1K	$+0.98_{[-0.91,2.86]}$	$+0.52_{[-1.76,2.80]}$	$-1.30_{[-3.45,0.91]}$

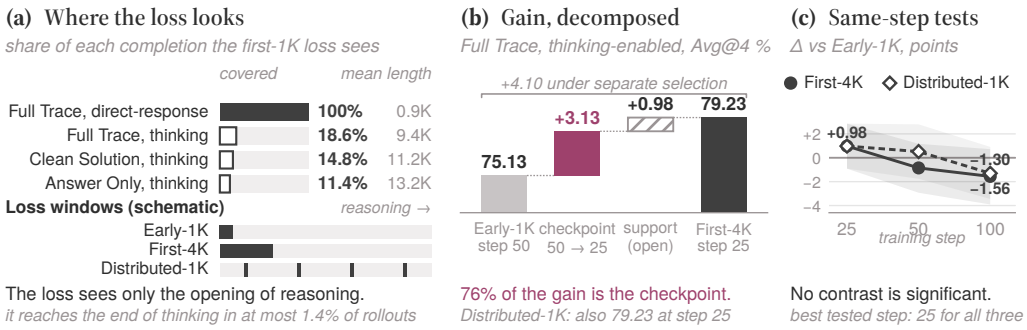


Figure 5: **Most of the apparent loss-window gain comes from checkpoint choice.** (a) Mean completion coverage of the inherited loss window and schematic alternatives. (b) Checkpoint and loss-window contributions to the development-selected gain. (c) Same-step contrasts with EARLY-1K, with 95% problem-cluster bootstrap intervals. The loss-window experiments use one training seed.

opening. For thinking-enabled FULL TRACE training, FIRST-4K extends the inherited EARLY-1K window to 4,096 tokens, while DISTRIBUTED-1K places four 256-token windows across the reasoning (Figure 5a).

Both alternatives exceed development-selected EARLY-1K by about four points, but stop at step 25 rather than its step 50. At common step 25 the apparent advantage falls below one point. The intervals for both alternatives' gains include zero at every matched checkpoint (Figure 5b–c), leaving stopping time as the main source of the apparent improvement (Dodge et al., 2020; Bouthillier et al., 2021).

6 Related Work

Learning from privileged context. On-policy distillation trains on the student's own responses, bringing teacher feedback to the states it actually visits (Agarwal et al., 2024). Conditioning the teacher on demonstrations (Shenfeld et al., 2026) or solutions (Zhao et al., 2026a) extends this idea to self-distillation, connecting it to learning using privileged information (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). This makes additional context a source of supervision without requiring a separate, stronger model. Its usefulness nevertheless depends on teacher–student compatibility and the training objective, now central questions in studies of on-policy distillation (Li et al., 2026a; Zhu et al., 2026).

What the reference contributes. Concurrent work is examining what privileged references contribute beyond distillation itself. Comparisons of reference types and reasoning modes find gains without target-specific solutions (Shrestha & Tessier, 2026; Ichihara et al., 2026), while structured guidance can outperform full traces (Zhao et al., 2026b). We construct AMPLE-MATH to hold answer availability fixed across six reasoning views. Using this reusable suite, we follow each reference's benefit beyond reference-free training across model families and checkpoints. Holding the references fixed, we also show that changing student training trajectories can reverse transfer in both families.

From teacher signals to student learning. Closer teacher imitation does not always improve generalization (Stanton et al., 2021), making it important to understand what the teacher's scores encourage. Work on OPSD identifies possible obstacles in preferences for the supplied solution (Harne et al., 2026), likelihood shifts without useful token credit (Nguyen et al., 2026), and suppressed reconsideration (Kaur et al., 2026). Complementary methods revise token targets, hidden-state supervision, or the teacher itself (Shen et al., 2026; Li et al., 2026c; Feng et al., 2026; Sun et al., 2026). We test whether changing these signals changes student behavior, pairing profiles on fixed and trained-student responses with marker-loss and loss-window interventions. Matching checkpoints also separates the effect of loss design from the benefit of stopping earlier.

7 Discussion and Conclusion

Successful distillation and a useful reference are different outcomes. Our answer-matched comparisons show that much of the improvement can occur without a reference, while a reference's additional benefit depends on the model and training stage. More complete solutions do not consistently yield greater gains. Replacing genuine references with unrelated ones can reduce accuracy, so reference-free gains do not make reference content irrelevant. Separating these effects makes reference selection a question of what helps the student beyond distillation itself.

The student's own responses are central to that choice because they determine the prefixes the teacher supervises. With the teacher and references unchanged, switching from short direct-response rollouts to long thinking-enabled rollouts under the same loss window turns gains into losses in both families. An asymmetry remains in the direct-response training configuration even without privileged information, since a thinking-enabled teacher supervises direct-response attempts. Learning from these scores may make existing reasoning capabilities more accessible through parameters shared by both inference modes. The concentration of gains on problems the base already sometimes solves is consistent with this interpretation.

Teacher profiles show how feedback varies across student responses and identify candidate changes to supervision. Relaxing loss at sampled correction markers barely changes their use, while matching checkpoints attributes most of the apparent benefit of broader loss coverage to stopping earlier. Changing the training signal does not necessarily produce the intended change in the student. Our evidence comes from short-run LoRA training on mathematics in two model families. These findings make reference design and student training a joint problem. They motivate testing whether reference content that adapts to the student's evolving attempts improves on a fixed representation throughout training. AMPLE-MATH provides a reusable setting for studying that interaction while holding answer information fixed. Combined with teacher profiles and matched interventions, it enables future OPSD research to test both what a reference adds and the explanations proposed for its effects. The central lesson is to judge privileged information by what it adds to the student, not by how complete a solution it gives the teacher.

AI USE STATEMENT

The authors determined the research questions, method, experimental design, analysis, and claims. Generative AI tools assisted with manuscript and software preparation. Their feedback helped the authors iteratively refine hypotheses, experimental design, mathematical formulation, and interpretation of results. Appendix B describes their use in generating and auditing the dataset's reasoning views. The authors edited and checked the manuscript, verified the reported results, and take responsibility for the work.

ETHICS STATEMENT

This work uses publicly released mathematical reasoning data and open model checkpoints and involves no human subjects or personal data. AMPLE-MATH preserves source provenance, and its release is subject to the licenses of its upstream data and models. As with other reasoning-model research, the trained models can generate incorrect or misleading outputs; our evaluation is limited to mathematical correctness and should not be interpreted as a general safety assessment.

REPRODUCIBILITY STATEMENT

Appendix A gives the objective, optimization, checkpoint and decoding settings, and statistical estimators. Appendix B describes dataset construction and split rules, and Appendices C to E provide the complete result tables. Code is available in the anonymous repository. AMPLE-MATH will be released with the camera-ready version (Appendix A.5).

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7527–7543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.468. URL https://aclanthology.org/2023.emnlp-main.468/.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305.

Yangyang Feng, Zhuoyan Feng, and Junlan Chen. Past: Privileged adaptation from complete student trajectories for on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.08726.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James

Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7xjoTuaNmN.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How pi-conditioned teachers break self-distillation, 2026. URL https://arxiv.org/abs/2608.04794.

HMMT. Harvard–MIT Mathematics Tournament: February 2025 problems and solutions. Official tournament archive, 2025. URL https://www.hmmt.org/www/archive/282.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? dissecting on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.09228.

Semih Kara and Oğuzhan Ersoy. The role of feedback alignment in self-distillation, 2026. URL https://arxiv.org/abs/2606.11173.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. URL https://arxiv.org/abs/2607.05184.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026a. doi: 10.48550/arXiv.2604.13016. URL https://arxiv.org/abs/2604.13016.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision, 2026b. URL https://arxiv.org/abs/2608.06296.

Yuhan Li, Mingxu Zhang, Dazhong Shen, and Ying Sun. Phf: Privileged hidden flow for on-policy self-distillation, 2026c. URL https://arxiv.org/abs/2606.29340.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Chen Lin, Kedi Chen, and Wei Zhang. Renio: Reweighting negative trajectory importance for llm on-policy distillation, 2026. URL https://arxiv.org/abs/2606.23104.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.03643.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

MAA. 2024 American Invitational Mathematics Examination: AIME I and II. Mathematical Association of America. Problems archived by Art of Problem Solving, 2024. URL https://artofproblemsolving.com/wiki/index.php/2024_AIME_I. AIME II: https://artofproblemsolving.com/wiki/index.php/2024_AIME_II.

MAA. 2025 American Invitational Mathematics Examination: AIME I and II. Mathematical Association of America. Problems archived by Art of Problem Solving, 2025. URL https://artofproblemsolving.com/wiki/index.php/2025_AIME_I. AIME II: https://artofproblemsolving.com/wiki/index.php/2025_AIME_II.

Xuan-Phi Nguyen, Zeyu Leo Liu, Yang Li, Shrey Pandit, Yiran Zhao, Anurag Koul, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.09263.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut, 1960. ISBN 9780598554512. URL https://books.google.com.sg/books?id=XnlevwEACAAJ.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified opsd: On-policy self-distillation without losing how to think, 2026. URL https://arxiv.org/abs/2607.02234.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Samyak Shrestha and Alexander Tessier. Rethinking privileged information in on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.18271.

Samuel Don Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2106.05945. URL https://proceedings.neurips.cc/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Zhuo Sun, Entong Li, Yanlong Zhao, Xiaoyuan Cheng, Wenxuan Yuan, Kaiyu Li, Che Liu, Huihang Liu, Harrison Bo Hua Zhu, and Li Zeng. Sr-opsd: Self-referenced on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.09745.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130. Advances in Neural Networks Research: IJCNN2009.

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Jpxfof0EaS.

Xuyang Zhao, Liting Zhang, Zichen Xu, Zhihu Wang, Xu Caiyue, Shiwan Zhao, and Qicheng Li. Is more privileged information better? from solution traces to problem-solving structure in self-distilled reasoning, 2026b. URL https://arxiv.org/abs/2608.01589.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes, 2026. URL https://arxiv.org/abs/2605.11182.

A Common Experimental Details

A.1 Implementation and optimization

Qwen3-1.7B uses a frozen backbone and LoRA adapters ($r = 64$, $\alpha = 128$, dropout 0) on the query, key, value, output, gate, up, and down projections. Training uses effective batch size 32, learning rate 5×10^{-6}, AdamW (Loshchilov & Hutter, 2019) with $\beta_2 = 0.999$, rollout and loss temperatures 1.1, and 100 optimizer steps. Checkpoints are saved at steps $0, 1, 2, 5, 10, 15, 20, 25, 50, 75, 100$.

Direct-response rollouts are capped at 1,024 tokens, all supervised, while thinking-enabled Qwen rollouts extend to 20,764 tokens with the same first-1,024-token loss support unless an intervention changes $I(y)$. We follow the released code's generalized divergence at $\beta = 0$ with a per-term clip of 0.05, rather than the $\beta = 0.5$ described in the paper (Zhao et al., 2026a).

SmolLM3-3B uses the same recipe with native thinking/no-thinking templates, a 24,536-token thinking-enabled rollout cap, and a 40,960-token training context. Its loss still covers the first 1,024 completion tokens. In both models the teacher is the frozen, thinking-enabled backbone with the student adapter disabled. Student rollouts and evaluation receive no PI. We evaluate the same trained checkpoints with thinking enabled and under direct-response generation.

A.2 Evaluation and effective correctness

In-domain evaluation draws four samples per problem on the frozen 384-problem test split at temperature 1.0, top-$p = 0.95$, and top-$k = 20$. Both families use a 32,768-token evaluation context and an initial 16,384-token generation budget. Every capped response is regenerated from the beginning with the same sampling seed, whether or not it already contains an answer. For a P-token prompt, the new-token limit is $\min(32{,}512, 32{,}768 - P - 256)$, reserving a 256-token margin. *Effective correctness* grades the extended-budget output when available, otherwise the initial response. External evaluation follows the released OPSD prompt and sampling protocol: 12 samples per problem, temperature 1.0, top-$p = 0.95$, and 38,912 generated tokens within a 40,960-token context. The external rescue trigger is inactive at this budget, so main-pass and rescue-aware scores coincide.

For saved-prefix comparisons, we decode the first 4,096, 8,192, or 16,384 token IDs from each initial response and apply the same answer grader. Shorter responses are graded in full. Only the full-evaluation endpoint uses the extended-budget output when available. These are grades of stored prefixes, not new generations asked to finish within a smaller budget.

Training-configuration comparisons primarily use common steps 25, 50, and 100, with development-selected checkpoints as secondary summaries. Development-selected checkpoints maximize Avg@2 on the 192-problem development split over steps 25, 50, 75, and 100, without test access. The selected steps for the seed-0 six-view comparison are 75, 50, 75, 75, 100, and 25 for Answer Only, Gist, Key Points, Clean Solution, Summary, and Full Trace, respectively.

A.3 Profile statistics

Writing the token shift as $\delta_{v,t}(a) = \log p^T_{v,t}(a) - \log p^S_t(a)$, correction pressure F_v averages $\delta_{v,t}(y_t)$ over sampled marker positions within each problem, then weights contributing problems equally. Temporal KL allocation reports the share of full-vocabulary $D_{\mathrm{KL}}(p^T_{v,t} \,\|\, p^S_t)$ in each quarter of the retained span. These statistics and correctness alignment are bootstrapped over problems.

For the correction-pressure diagnostic, token pieces are lowercased after replacing tokenizer space markers with spaces, then matched by substring against *wait*, *but*, *however*, *maybe*, *actually*, *recheck*, *check*, *verify*, *reconsider*, *mistake*, and *instead*. This lexical rule identifies sampled marker tokens rather than all semantically corrective continuations.

A.4 Statistical estimators

The problem is the inferential unit. Avg@n first averages the n sampled correctness indicators within problem and then averages across problems. Paired effects are differences of per-problem averages. We construct 95% percentile intervals from 10,000 problem-cluster bootstrap resamples, reusing one resampled problem-index array across compared conditions. For balanced in-domain tests, resampling is stratified by the construction strata. Pooled external intervals resample all 90 problems together, without benchmark stratification. We distinguish paired intervals excluding zero from effects surviving Holm correction (Holm, 1979) across a family of comparisons. Intervals including zero do not establish equivalence. Reported p-values are two-sided bootstrap tail probabilities, twice the smaller of the resampled fractions at or below and at or above zero, floored at $1/10{,}000$.

Seed-aware aggregates first average the paired effects within problem across the observed seeds, then bootstrap the resulting problem vector. These intervals describe problem-sampling uncertainty

for the observed seeds. Training-seed variability is reported separately through per-seed effects and seed-level t intervals. Paired seed-level checks use the four matched ANSWER ONLY-minus-FULL TRACE differences, while the CLEAN SOLUTION-minus-reference-free check uses a Welch interval over three seed means per configuration.

We apply Holm's procedure to the six view-versus-base tests, the six in-domain view-versus-reference-free comparisons, and the two external comparisons of ANSWER ONLY and FULL TRACE with the reference-free control. For interventions, we correct within each control family (four marker-mask, two other-problem, three trace-opening, and two prompt-template comparisons), across all eleven controls, and across the six loss-support contrasts.

A.5 RELEASE MATERIALS AND COMPUTATION

Code is available in the anonymous repository. AMPLE-MATH will be released with the camera-ready version. Row-level evaluation outputs are available on request. We estimate total project compute at approximately 1,000 H100 GPU-hours, including training, evaluation, teacher profiling, dataset construction, and setup.

B AMPLE-MATH CONSTRUCTION AND AUDITING

B.1 SOURCE, FILTERING, AND THE FROZEN POPULATION

The source is the `metadata` configuration of OpenThoughts-114k (Guha et al., 2026): one problem, one verified answer, and one natural reasoning trace per row. We required a gradable answer and a source trace whose final answer agrees with it, collapsed within-corpus near duplicates, excluded source traces above 16,384 tokens, and screened for exact and semantic overlap with MATH-500 (Lightman et al., 2024), AIME 2024, and JEEBench (Arora et al., 2023). We then selected 5,480 eligible questions for the frozen pool. Construction lengths use the fast Llama-3.1-8B-Instruct tokenizer without special tokens. A question enters the complete resource only if all six views exist, every generated view passes surface validation and source-fidelity review, and body lengths satisfy ANSWER ONLY $<$ GIST $<$ KEY POINTS $<$ SUMMARY $<$ FULL TRACE. CLEAN SOLUTION is off this length spine and is not constrained relative to the generated views. In total, 5,319 do, giving $5{,}319 \times 6 = 31{,}914$ supervision records and leaving 161 excluded questions.

A word-shingle scan of the 90 external evaluation problems (AIME 2024, AIME 2025, and HMMT February 2025) against the 5,319 complete resource problems found no exact or near duplicates. The largest 8-gram containment of any benchmark problem was 0.25, from shared contest phrasing. Because AMPLE-MATH and the released OPSD training data draw on the same upstream source, 202 of the 384 test problems (52.6%) also appear in that released data. Only the original-data implementation check trains on those data, and it is evaluated on the external benchmarks rather than the overlapping in-domain test set. Every representation and rollout comparison uses the local resource with disjoint train, development, and test splits.

B.2 THE SIX ANSWER-MATCHED VIEWS

ANSWER ONLY has no reasoning body, while CLEAN SOLUTION and FULL TRACE use the source's polished solution and complete trace. Qwen3.6-35B-A3B-FP8 compresses the intact trace into a short paragraph for GIST, ordered steps for KEY POINTS, and a narrative for SUMMARY, without a separate verified-answer field (Qwen Team, 2026). Up to five reproducible, seeded attempts use temperatures 0.0, 0.3, 0.6, 0.9, and 1.2. Candidates undergo surface validation for completion, form, and prompt leakage, then source-fidelity review for unsupported claims, backward-constructed reasoning, and confidence upgrades. Every accepted view passes both stages. Fidelity review preserves the source's reasoning and uncertainty rather than certifying or repairing its proof. Generation and review use the same model under different prompts, so the review may share the generator's blind spots.

All views share the canonical answer but vary the reasoning's representation, wording, and detail. Records render as `<think>` body `</think><answer>` answer `</answer>`, with exactly one boxed answer in the answer section. The aligned example shows how the six views express the same solution.

An aligned example.

Shared problem What is the value of $6 \times (5 - 2) + 4$?
Canonical answer 22

PI view	Reasoning body (excerpts)
ANSWER ONLY	*Empty reasoning body.*
GIST	Apply order of operations: evaluate parentheses, then multiply, then add.
KEY POINTS	Key steps: – Apply order of operations, starting with parentheses [. . .] – Complete the final addition: 18 plus 4 equals 22
CLEAN SOLUTION	The value of $6 \times (5 - 2) + 4$ is calculated as follows: 1. **Parentheses first**: $5 - 2 = 3$. [. . .] 3. **Add**: $18 + 4 = 22$. **Final Answer:** $\boxed{22}$
SUMMARY	Summary: I begin by identifying the expression as a standard arithmetic problem requiring strict adherence to the order of operations, [. . .] leading to the final value of twenty-two.
FULL TRACE	Okay, let's see. I need to figure out the value of 6 times (5 minus 2) plus 4. Hmm, order of operations, right? I remember learning PEMDAS–Parentheses, Exponents, Multiplication and Division, Addition and Subtraction. [. . .] Wait, let me double-check to make sure I didn't skip any steps. [. . .] Yep, 22 should be the correct answer.

Original wording is preserved, with Markdown and math typeset. Omissions are marked [. . .].

B.3 QWEN AND SMOLLM3 SPLIT CONSTRUCTION

We drew the train, development, and test sets from 3,399 candidates remaining after 1,920 prior exclusions from the 5,319 complete problems, keeping the profiling set disjoint from all three splits. For Qwen3-1.7B, four problem-only direct-response samples from the frozen base define high-success ($p \geq 0.75$), intermediate ($p \in \{0.25, 0.5\}$), and zero-success ($p = 0$) bands. A zero-success problem is eligible for the third band only when at least one structured PI teacher generation is correct. The final split contains 512/64/128 problems per band in train/dev/test, for totals $1{,}536/192/384$.

SmolLM3 uses the same development and test IDs but a training split rebuilt from its own direct-response outcomes. The same structured-PI recoverability rule permits 512 training problems per band. On the shared test IDs, SmolLM3 has 129 high-success, 97 intermediate, and 158 zero-success problems. Overall performance on all 384 is primary, with these unequal bands used for secondary analysis.

B.4 DIFFICULTY AND VIEW LENGTHS

Dataset difficulty is distinct from the student-relative outcomes used for the experimental splits. Three evaluator models (Qwen2.5-1.5B-Instruct, Llama-3.1-8B-Instruct, and Qwen3.6-35B-A3B-FP8, Qwen et al., 2025; Grattafiori et al., 2024; Qwen Team, 2026) answered each of the 5,480 frozen questions from the statement alone, with 8 to 16 attempts each. A binomial Rasch model (Rasch, 1960) jointly fits evaluator abilities and question difficulties to each evaluator's correct and actually graded counts, rather than the nominal number of attempts. A zero-centered L_2 penalty of 0.5 keeps all-correct and all-incorrect items finite. Difficulty measures final-answer success under a fixed decoding and grading protocol, not proof quality. Figure 6 shows the difficulty distribution and its relation to source-trace length. Figure 7 compares reasoning-body lengths across views.

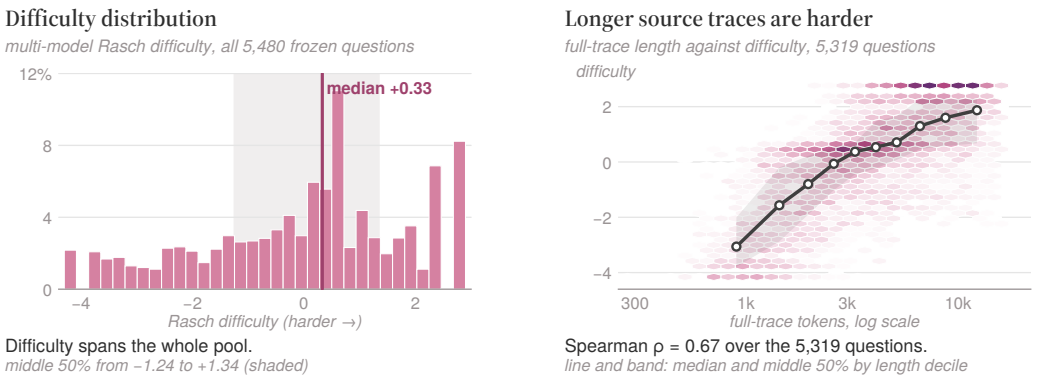


Figure 6: **Difficulty and source-trace length in AMPLE-MATH.** Left: multi-model Rasch difficulty over the 5,480-question frozen pool. Right: full-trace token length versus difficulty for the 5,319 complete six-view problems. Hexagon shading counts problems, and the line and band show median difficulty and its interquartile range within source-trace length deciles.

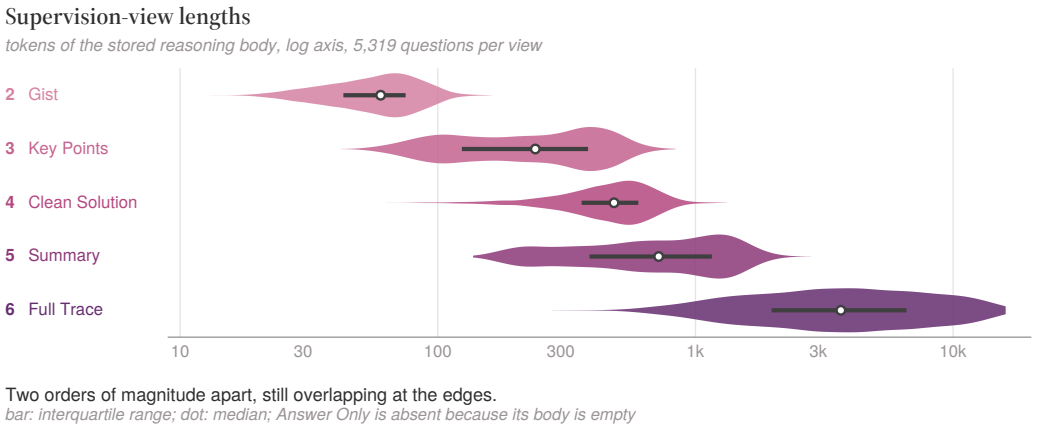


Figure 7: **Reasoning-body lengths across the five nonempty views.** Token-length distributions of the stored reasoning bodies over the 5,319 complete questions, on a logarithmic axis, with the interquartile range and median marked; ANSWER ONLY is absent because its body is empty by construction. These lengths use the Llama-3.1-8B-Instruct construction tokenizer. Main-text lengths measure the rendered privileged reference under the Qwen3 tokenizer.

C STUDENT-TRANSFER EVIDENCE

The evidence follows the main paper's distinction between distillation gains and reference contributions, then checks where the gains occur and how they depend on training. Secondary direct-response and saved-prefix analyses appear at the end. All evaluation is unprivileged and follows the common protocol in Appendix A.

C.1 REFERENCE CONTRIBUTION IN-DOMAIN AND EXTERNALLY

The reference-free teacher is the frozen thinking-enabled backbone scoring the student's prefix without PI. The wrong-answer control uses ANSWER ONLY with the canonical answer replaced by another problem's answer in the same format. No training problem retains its own answer. Both controls use three training seeds. At step 100, the three-seed CLEAN SOLUTION comparison gives $+1.30\,[0.20, 2.41]$ points over the three-seed reference-free control. This small unadjusted advantage does not survive Holm correction across the six-view family.

Tables 3 and 4 give the Qwen controls and all six views' contrasts against each control. In this six-view comparison, view rows use seed 0 while each control averages its three seeds. Only CLEAN SOLUTION's unadjusted interval over the reference-free student lies above zero (raw $p = 0.024$, Holm-adjusted 0.15).

Table 3: **Teacher controls.** Δ Avg@4 versus the frozen base with paired 95% intervals at common checkpoints. Three-seed rows average per-problem effects over seeds 0–2.

Control	Step 25	Step 50	Step 100
Reference-free (seed 0)	$+1.63_{[0.00,3.26]}$	$+3.26_{[1.50,5.08]}$	$+1.76_{[-0.20,3.71]}$
Reference-free (3 seeds)	–	$+2.89_{[1.39,4.45]}$	$+1.80_{[0.20,3.47]}$
ANSWER ONLY, wrong answer (seed 0)	$+1.30_{[-0.52,3.12]}$	$+2.28_{[0.59,3.97]}$	$+0.98_{[-1.04,2.99]}$
ANSWER ONLY, wrong answer (3 seeds)	–	$+3.06_{[1.58,4.60]}$	$+1.67_{[0.07,3.32]}$

Table 4: **Views against the controls.** Each seed-0 view minus each three-seed control at step 100, in Avg@4 points with unadjusted paired 95% intervals.

View	minus reference-free	minus wrong-answer ANSWER ONLY
ANSWER ONLY	$+0.35_{[-1.04,1.71]}$	$+0.48_{[-0.85,1.76]}$
GIST	$+0.35_{[-1.13,1.80]}$	$+0.48_{[-0.93,1.87]}$
KEY POINTS	$+1.00_{[-0.48,2.45]}$	$+1.13_{[-0.37,2.58]}$
CLEAN SOLUTION	$+1.78_{[0.24,3.30]}$	$+1.91_{[0.52,3.30]}$
SUMMARY	$+0.74_{[-0.69,2.17]}$	$+0.87_{[-0.52,2.24]}$
FULL TRACE	$+0.61_{[-0.87,2.11]}$	$+0.74_{[-0.67,2.15]}$

Table 5 gives the external counterpart to main Table 1, comparing each PI-trained student with the reference-free student rather than the frozen base. No external view-minus-control interval lies entirely above zero. The unadjusted FULL TRACE-minus-control interval lies below zero, but the contrast does not survive Holm correction over the two extreme-view tests.

Table 5: **External view-minus-control comparisons.** Step-50 Avg@12 differences against the reference-free student, with unadjusted paired 95% intervals over 90 problems. ANSWER ONLY and FULL TRACE average four training seeds, the control three, and the other views use one. Absolute effects and benchmark breakdowns are in Table 1.

View	Difference vs. reference-free student
ANSWER ONLY	$-0.35_{[-1.64,1.00]}$
GIST	$+0.28_{[-1.94,2.47]}$
KEY POINTS	$-0.46_{[-2.96,1.98]}$
CLEAN SOLUTION	$+0.83_{[-1.05,2.78]}$
SUMMARY	$-0.46_{[-2.53,1.54]}$
FULL TRACE	$-1.78_{[-3.43,-0.21]}$

The SmolLM3 comparison keeps steps 50 and 100 together with per-seed gains and completion diagnostics (Table 6). FULL TRACE improves on the reference-free students with thinking enabled at both checkpoints, but by step 100 all three configurations fall below the frozen base. The table also retains direct-response outcomes, where the late deterioration is especially pronounced. Even after re-generation, the reference-free and ANSWER ONLY students' direct-response no-answer rates rise from 20.2% and 22.0% at step 50 to 42.0% and 39.4% at step 100. Table 6 separates final-output accuracy from initial-pass completion diagnostics. For thinking-enabled FULL TRACE at step 50, re-generation reduces the no-answer rate from 19.7% to 1.3%.

Table 6: **SmolLM3-3B reference comparisons across checkpoints and evaluation modes.** All trained configurations use direct-response training and three seeds. **A:** Final-output Avg@4 gains over the frozen base, in percentage points with 95% problem-cluster bootstrap intervals. Parenthesized triplets give gains for seeds 0, 1, and 2. **B:** Paired FULL TRACE-minus-reference-free contrasts with the same interval convention. **C:** Initial-pass mean lengths, 16,384-token cap rates, and no-answer rates, before re-generation. No answer means no extractable \boxed{...} answer. Extraction uses the last boxed answer in the text. Diagnostics pool 4,608 responses per trained configuration. Frozen-base Avg@4 is 82.16% with thinking enabled and 40.23% under direct-response evaluation.

A. Accuracy gains versus the frozen base

Student	Thinking-enabled	Direct-response
Step 50		
Reference-free	$+0.59_{[-1.06,2.26]}$ $(+0.20, +0.46, +1.11)$	$+13.24_{[9.96,16.49]}$ $(+12.76, +11.65, +15.30)$
■ ANSWER ONLY	$+0.87_{[-0.74,2.50]}$ $(+0.46, +1.37, +0.78)$	$+11.91_{[8.57,15.34]}$ $(+12.11, +12.50, +11.13)$
■ FULL TRACE	$+2.58_{[1.06,4.14]}$ $(+2.93, +2.41, +2.41)$	$+2.60_{[0.46,4.77]}$ $(+1.56, +3.58, +2.67)$
Step 100		
Reference-free	$-5.25_{[-7.20,-3.32]}$ $(-4.75, -6.25, -4.75)$	$-10.39_{[-13.15,-7.68]}$ $(-11.72, -9.70, -9.77)$
■ ANSWER ONLY	$-5.03_{[-6.94,-3.17]}$ $(-5.14, -4.88, -5.08)$	$-10.09_{[-12.85,-7.38]}$ $(-11.78, -12.76, -5.73)$
■ FULL TRACE	$-2.19_{[-4.08,-0.37]}$ $(-3.97, -1.24, -1.37)$	$+2.76_{[0.39,5.14]}$ $(-0.72, +5.21, +3.78)$

B. Full Trace minus reference-free

Step 50	$+2.00_{[0.82,3.17]}$	$-10.63_{[-13.19,-8.07]}$
Step 100	$+3.06_{[1.71,4.41]}$	$+13.15_{[10.94,15.39]}$

C. Initial-pass length and completion diagnostics

		Thinking-enabled			Direct-response		
Student	**Step**	**Tokens**	**Cap**	**No answer**	**Tokens**	**Cap**	**No answer**
Frozen base	0	7,492	12.9%	10.9%	833	0.3%	0.8%
Reference-free	50	9,078	19.1%	18.3%	5,760	2.9%	21.3%
■ ANSWER ONLY	50	9,485	20.5%	19.0%	5,959	3.0%	23.4%
■ FULL TRACE	50	9,746	22.6%	19.7%	1,809	0.6%	3.3%
Reference-free	100	8,700	15.1%	17.1%	11,153	42.8%	52.5%
■ ANSWER ONLY	100	8,866	16.2%	17.1%	10,875	43.2%	49.8%
■ FULL TRACE	100	9,841	23.4%	20.8%	6,629	25.5%	23.5%

C.2 WHERE THE GAINS OCCUR

Table 7 reports the step-100 multi-seed Qwen gains within the three frozen split-construction groups, matching Figure 2c. The largest gains occur in the zero-correct direct-response group, where the frozen base nevertheless answers 61.91% of samples correctly with thinking enabled. Within this group, CLEAN SOLUTION exceeds reference-free training by 3.06 $[0.46, 5.79]$ points, its largest additional gain among the three groups. The intervals describe unadjusted within-group comparisons.

Validation on the original training data. The implementation check retains the original OPSD training data and settings (Zhao et al., 2026a), with direct-response student rollouts capped at 1,024 tokens and a thinking-enabled teacher. Unlike the preceding AMPLE-MATH comparisons, this check evaluates a student trained on the original data. Table 8 reports step-50 Avg@12 on the external benchmarks. Pooled Majority@12 rises by 7.78 points while Pass@12 falls by 7.78 points.

Grouping the 90 external problems by the frozen base's 12 thinking-enabled samples gives 27 never-solved, 49 intermittently solved, and 14 always-solved problems. Their Avg@12 changes are

Table 7: **Multi-seed gains by split-construction group.** Qwen3-1.7B Avg@4 changes in percentage points at step 100 after direct-response training, with thinking-enabled evaluation. Groups contain 128 test problems each and are defined by correct answers among four frozen-base direct-response samples. Estimates average three training seeds for reference-free and CLEAN SOLUTION, and four for ANSWER ONLY and FULL TRACE; subscripts give unadjusted 95% paired problem-bootstrap intervals (10,000 resamples within each group). The final row gives frozen-base thinking-enabled accuracy, not a gain.

Configuration	Zero correct 0/4	Intermediate 1–2/4	High success 3–4/4
Gain over frozen base			
Reference-free	$+3.97_{[0.39,7.62]}$	$+0.98_{[-1.82,3.91]}$	$+0.46_{[-1.17,2.28]}$
ANSWER ONLY	$+4.69_{[1.32,8.25]}$	$+2.34_{[-0.24,4.98]}$	$+0.20_{[-1.76,2.25]}$
CLEAN SOLUTION	$+7.03_{[3.52,10.74]}$	$+1.89_{[-0.72,4.62]}$	$+0.39_{[-1.43,2.34]}$
FULL TRACE	$+5.66_{[2.59,8.84]}$	$+1.51_{[-1.12,4.25]}$	$-0.05_{[-1.76,1.81]}$
Paired gain over reference-free			
ANSWER ONLY	$+0.72_{[-1.51,2.99]}$	$+1.37_{[-0.26,3.03]}$	$-0.26_{[-1.20,0.62]}$
CLEAN SOLUTION	$+3.06_{[0.46,5.79]}$	$+0.91_{[-0.85,2.60]}$	$-0.07_{[-0.98,0.78]}$
FULL TRACE	$+1.69_{[-0.62,4.00]}$	$+0.54_{[-1.46,2.54]}$	$-0.50_{[-1.33,0.31]}$
Frozen-base accuracy	61.91%	83.98%	94.73%

Table 8: **Validation on the original OPSD training data.** Thinking-enabled external evaluation at step 50. Values are Avg@12 percentages, with 95% intervals for the changes from the frozen base.

	AIME24	AIME25	HMMT25	Pooled
Frozen base	47.22	39.17	23.89	36.76
OPSD step 50	54.72	42.78	29.44	42.31
Difference	+7.50	+3.61	+5.56	+5.56
95% CI	[2.22, 12.78]	[−2.50, 10.56]	[−0.56, 12.50]	[2.22, 9.17]

$+0.62$, $+10.03$, and -0.60 points, respectively. This descriptive breakdown locates the gain among intermittently solved problems. These groups differ from the direct-response construction strata in AMPLE-MATH.

C.3 TRAINING-SEED AND CHECKPOINT CHECKS

The Qwen extreme views ANSWER ONLY and FULL TRACE each use four seeds under the same data, optimizer, and schedule. Table 9 reports their paired effects and between-seed ranges, using step 100 in domain and step 50 externally.

Table 10 compares problem-bootstrap and seed-level uncertainty for the principal step-100 contrasts. Both checks support a small unadjusted CLEAN SOLUTION advantage over reference-free training and an evaluation-mode interaction for ANSWER ONLY versus FULL TRACE. The seed-level check conditions on the test set, while the problem bootstrap conditions on the observed seeds.

At step 100, the three-seed CLEAN SOLUTION contrasts against the four-seed ANSWER ONLY and FULL TRACE averages are $0.69\,[-0.22, 1.65]$ and $0.73\,[-0.17, 1.65]$ points, respectively. At step 50, the two available CLEAN SOLUTION seeds (0 and 2) give a $1.05\,[0.07, 2.06]$-point advantage over the three-seed reference-free control. Under development selection, the ANSWER ONLY-minus-FULL TRACE contrast over seeds 0–2 is $-0.24\,[-1.28, 0.76]$.

C.4 TRAINING-TRAJECTORY COMPARISONS AND COMPLETION CHECKS

Table 11 gives Qwen accuracy levels and paired training-configuration gaps at common checkpoints. All nine gaps are negative, and only CLEAN SOLUTION at step 25 has an interval that includes zero.

Table 9: **Training-seed robustness of the extreme-view contrast.** Changes versus the frozen base, with paired 95% problem-bootstrap intervals. In-domain evaluation uses step-100 Avg@4 on 384 problems. External evaluation uses step-50 Avg@12 on 90 problems. Seed-aware estimates average per-problem effects over four seeds before resampling.

	In domain		External	
	■ ANSWER ONLY	■ FULL TRACE	■ ANSWER ONLY	■ FULL TRACE
Seed 0	$+2.15_{[0.26,4.04]}$	$+2.41_{[0.65,4.17]}$	$+3.70_{[0.93,6.67]}$	$+2.87_{[0.00,5.93]}$
Seed 1	$+1.89_{[0.07,3.78]}$	$+3.06_{[1.30,4.88]}$	$+3.70_{[0.74,6.76]}$	$+1.39_{[-1.39,4.26]}$
Seed 2	$+3.06_{[1.24,4.95]}$	$+2.80_{[0.98,4.62]}$	$+3.33_{[0.37,6.57]}$	$+1.57_{[-1.57,4.72]}$
Seed 3	$+2.54_{[0.72,4.30]}$	$+1.24_{[-0.65,3.06]}$	$+4.17_{[1.30,7.22]}$	$+3.33_{[0.56,6.11]}$
Seed-aware mean	$+2.41_{[0.86,3.99]}$	$+2.38_{[0.91,3.86]}$	$+3.73_{[1.27,6.41]}$	$+2.29_{[-0.16,4.88]}$
Between-seed range	1.17	1.82	0.83	1.94
ANSWER ONLY minus FULL TRACE	$+0.03_{[-0.91,0.98]}$		$+1.44_{[-0.23,3.17]}$	

Table 10: **Problem-level and training-seed uncertainty give consistent readings.** Qwen3-1.7B at step 100, with Avg@4 contrasts in percentage points. Problem intervals average observed seeds before bootstrapping problems. Seed-level intervals use four paired differences (t, 3 degrees of freedom), except CLEAN SOLUTION versus reference-free, which uses three seeds each (Welch, approximately 2.3 degrees of freedom). All intervals are 95% and unadjusted.

Contrast	Evaluation	Problem bootstrap	Seed-level t
CLEAN SOLUTION − reference-free	Thinking-enabled	$+1.30_{[0.20,2.41]}$	$+1.30_{[0.27,2.33]}$
ANSWER ONLY − FULL TRACE	Thinking-enabled	$+0.03_{[-0.91,0.98]}$	$+0.03_{[-1.61,1.68]}$
ANSWER ONLY − FULL TRACE	Direct-response	$+5.32_{[3.47,7.16]}$	$+5.32_{[1.9,8.7]}$
ANSWER ONLY − FULL TRACE	Mode interaction	$+5.29_{[3.26,7.32]}$	$+5.29_{[2.62,7.96]}$

Table 11: **Fixed-checkpoint training-trajectory comparison.** Qwen Avg@4 under unprivileged, thinking-enabled evaluation. DR and TH denote direct-response and thinking-enabled training rollouts. Gaps are TH minus DR, in points with paired 95% problem-bootstrap intervals. The frozen base scores 80.21%.

View	Step	DR	TH	Gap [95% CI]
■ ANSWER ONLY	25	82.55	79.17	$-3.39_{[-5.40,-1.43]}$
	50	83.59	75.20	$-8.40_{[-11.00,-5.92]}$
	100	82.36	77.28	$-5.08_{[-7.23,-2.99]}$
■ CLEAN SOLUTION	25	82.03	81.05	$-0.98_{[-2.80,0.85]}$
	50	84.64	77.54	$-7.10_{[-9.38,-4.95]}$
	100	83.79	77.28	$-6.51_{[-8.79,-4.23]}$
■ FULL TRACE	25	82.55	78.26	$-4.30_{[-6.25,-2.41]}$
	50	82.42	75.13	$-7.29_{[-9.44,-5.14]}$
	100	82.62	72.79	$-9.83_{[-11.98,-7.68]}$
Pooled	25	–	–	$-2.89_{[-4.08,-1.71]}$
	50	83.55	75.95	$-7.60_{[-9.31,-5.95]}$
	100	82.92	75.78	$-7.14_{[-8.66,-5.64]}$

The FULL TRACE comparison is repeated on SmolLM3 with two matched training seeds (Table 12). At step 50, mean direct-response-trained and thinking-enabled-trained accuracies are 84.83% and 77.99%, against a base of 82.16%. FULL TRACE's final-output cap rate is at most 3.9% in both seeds.

Table 12: **SmolLM3 Full Trace replication.** Thinking-enabled-trained minus direct-response-trained Avg@4 for two matched training seeds, in points with paired 95% intervals.

	Step 25	Step 50	Step 100
Seed 0	$-3.78_{[-5.92,-1.56]}$	$-6.84_{[-8.98,-4.69]}$	$-6.64_{[-8.98,-4.23]}$
Seed 1	$-3.12_{[-5.01,-1.24]}$	$-6.84_{[-8.85,-4.82]}$	$-9.83_{[-12.11,-7.55]}$
Seed-aware mean	$-3.45_{[-4.98,-1.89]}$	$-6.84_{[-8.50,-5.18]}$	$-8.24_{[-10.16,-6.28]}$
Raw seed spread	0.65	0.00	3.19

The Qwen completion checks likewise distinguish ANSWER ONLY from the denser views. At development-selected checkpoints, thinking-enabled ANSWER ONLY training produces longer final outputs and more capping than direct-response training, yet its accuracy penalty is smaller than FULL TRACE's. The CLEAN SOLUTION and FULL TRACE thinking-enabled students instead produce shorter responses while also losing accuracy. The training-configuration gap thus persists after the extended-budget pass and for FULL TRACE, where capping is modest.

Table 13 reports the step-50 FULL TRACE training-configuration contrast by external benchmark. The penalty's interval lies below zero on AIME 2024 and AIME 2025, but includes zero on HMMT 2025. Against the base, the pooled effects are $-6.57\,[-9.81, -3.43]$ points for thinking-enabled training and $+2.87\,[0.00, 5.93]$ for direct-response training, whose interval touches zero.

Table 13: **External Full Trace pair.** Qwen Full Trace at the common step 50, with Avg@12 percentages and Δ denoting thinking-enabled-trained minus direct-response-trained.

	AIME24	AIME25	HMMT25	Pooled
Frozen base	47.22	39.17	23.89	36.76
Direct-response trained	53.89	40.56	24.44	39.63
Thinking-enabled trained	40.83	27.78	21.94	30.19
Δ	-13.06	-12.78	-2.50	-9.44
95% CI	$[-20.56, -5.83]$	$[-19.72, -6.39]$	$[-6.39, 1.11]$	$[-13.24, -5.83]$

The thinking-enabled-trained student reaches the 38,912-token cap in 79 of 1,080 external samples, versus two for the direct-response-trained student. Capped samples are scored as generated. Excluding them leaves a gap of $-8.16\,[-12.16, -4.25]$ points. Scoring every capped sample as correct, although none was, gives $-2.13\,[-6.85, +2.69]$, whose interval includes zero.

C.5 EVALUATION-MODE AND SAVED-PREFIX SENSITIVITY

Figure 8 evaluates the same checkpoints with thinking enabled and disabled, alongside the step-50 saved-prefix comparison. Table 14 gives the four-seed Qwen ANSWER ONLY-minus-FULL TRACE contrast in each mode and its paired interaction at steps 50 and 100. For SmolLM3 at step 50, the corresponding interaction is $11.02\,[8.14, 13.95]$ points (Table 6), defined as the direct-response contrast minus the thinking-enabled contrast.

Table 15 gives the corresponding step-100 comparison when stored Qwen direct responses are graded at progressively longer prefixes. Both ANSWER ONLY and reference-free students fall below FULL TRACE at 4K, then exceed it at 8K and beyond. The reference-free gain over the base has an interval above zero only at 16K and full evaluation, while FULL TRACE's interval includes zero at every cutoff.

D ADDITIONAL PRIVILEGE-PROFILE EVIDENCE

D.1 COVERAGE AND SAME-PREFIX MATCHING

The thinking-enabled profile uses 512 problems and four fixed no-PI trajectories per problem. Across 18,467,995 distinct student token positions, six PI contexts plus the no-PI context produce 129,275,965 scored token–condition evaluations. Every condition shares the span fitting all teacher contexts,

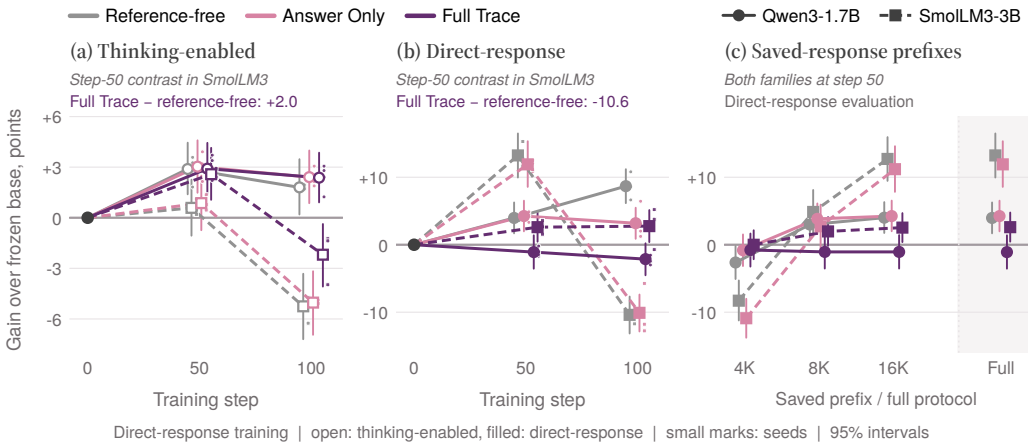


Figure 8: **Student gains under both evaluation modes and saved-prefix grading.** (a–b) Gains over the frozen base at matched checkpoints, evaluated with thinking enabled (a) and directly (b). (c) Step-50 direct-response prefixes from the initial generation pass. Full evaluation uses the final graded outputs. Means and 95% problem-cluster intervals use 384 problems and three seeds, except four for Qwen ANSWER ONLY/FULL TRACE. Step 0 is the shared frozen base.

Table 14: **Evaluation-mode interaction by training seed.** ANSWER ONLY minus FULL TRACE in Avg@4 percentage points under each evaluation mode, with interaction defined as the direct-response minus thinking-enabled contrast. Intervals are paired 95% problem-bootstrap intervals. Individual seed rows are at step 100. Seed-aware rows average per-problem effects over four seeds before resampling.

	Direct-response	**Thinking-enabled**	**Interaction**
seed 0	$+3.71_{[0.98,6.45]}$	$-0.26_{[-1.95,1.43]}$	$+3.97_{[0.65,7.23]}$
seed 1	$+3.52_{[0.85,6.25]}$	$-1.17_{[-2.74,0.39]}$	$+4.69_{[1.56,7.75]}$
seed 2	$+8.01_{[5.01,11.00]}$	$+0.26_{[-1.56,2.08]}$	$+7.75_{[4.23,11.13]}$
seed 3	$+6.05_{[3.19,8.98]}$	$+1.30_{[-0.20,2.86]}$	$+4.75_{[1.56,8.01]}$
Step 100, seed-aware	$+5.32_{[3.47,7.16]}$	$+0.03_{[-0.91,0.98]}$	$+5.29_{[3.26,7.32]}$
Step 50, seed-aware	$+5.34_{[3.69,7.00]}$	$+0.11_{[-0.76,0.99]}$	$+5.22_{[3.42,7.05]}$

Table 15: **Direct-response contrasts change when stored responses are graded at shorter lengths.** Qwen3-1.7B at step 100, averaging four ANSWER ONLY and FULL TRACE training seeds and three reference-free seeds. Entries are Avg@4 differences in points with paired 95% intervals from 10,000 shared problem-cluster bootstrap resamples. The first three rows grade decoded prefixes of 4,096, 8,192, and 16,384 main-pass tokens, without the second pass. The final row uses the full evaluation protocol.

Prefix	ANSWER ONLY − FULL TRACE	Reference-free − FULL TRACE	Reference-free − base	FULL TRACE − base
4K	$-3.01_{[-4.85,-1.22]}$	$-3.99_{[-6.26,-1.78]}$	$-5.95_{[-8.53,-3.39]}$	$-1.95_{[-4.33,0.36]}$
8K	$+3.29_{[1.46,5.06]}$	$+3.86_{[1.75,5.98]}$	$+1.78_{[-0.74,4.25]}$	$-2.08_{[-4.46,0.26]}$
16K	$+5.09_{[3.26,6.92]}$	$+9.51_{[7.41,11.62]}$	$+7.42_{[4.93,9.85]}$	$-2.08_{[-4.46,0.26]}$
Full	$+5.32_{[3.47,7.16]}$	$+10.81_{[8.60,13.00]}$	$+8.68_{[6.16,11.20]}$	$-2.13_{[-4.52,0.21]}$

capped at 1,024 completion tokens for direct-response profiles. Both profiles require at least 512 matched tokens. This excludes 14.3% of direct-response completions, leaving 1,756 trajectories from 466 problems. The teacher uses the prompt below with thinking enabled in both profiles, and omitted completions are not imputed. Correctness alignment requires both correct and incorrect scored completions, leaving 116 eligible problems for thinking-enabled profiles and 200 for direct-

response profiles. Their paired intersection contains 48 problems, so paired cross-mode estimates use a different population from the separate mode averages.

The student uses the shared block below, and the teacher appends the teacher-only block. Lines are wrapped for display. Both use the Qwen chat template with a generation prompt. The student-generation thinking flag selects the prefix distribution, while teacher scoring keeps thinking enabled.

Profiling prompt

Shared by student and teacher

```
Solve the following math problem.  Reason step by step, and put the
final answer in \boxed{}.

Problem:
{question}
```

Teacher-only addition

```
Privileged reference available only to the teacher:
<reference>
{condition_target}
</reference>

Use the reference to help solve or evaluate the problem, but do not
mention that a reference was provided.
```

D.2 Alignment on frozen and trained-student responses

For the trained-student profile, we generated responses to the 512 profiling problems from each seed-0 direct-response-trained student at step 50 and scored them with the same frozen teacher under all seven contexts. Prompts and generation seeds match the pre-training procedure. Table 16 compares the frozen-base profiles with alignment on each trained student's own responses. The responses and the set of problems with both correct and incorrect completions can therefore change after training; the teacher remains fixed. Five views have much smaller estimates after training, while Full Trace retains about three-quarters of its initial value.

Table 16: **Correctness alignment across student trajectories.** C_v in nats per token with 95% problem-bootstrap intervals. The teacher is thinking-enabled under both student response modes. The final column scores each trained student's own direct-response completions at step 50.

View	Frozen base thinking enabled	Frozen base direct response	Own step-50 states direct response
Answer Only	$+0.0065_{[0.0041,0.0091]}$	$+0.0041_{[0.0028,0.0054]}$	$+0.0013_{[0.0002,0.0024]}$
Gist	$+0.0042_{[0.0015,0.0072]}$	$+0.0067_{[0.0045,0.0090]}$	$+0.0024_{[0.0001,0.0046]}$
Key Points	$+0.0019_{[-0.0011,0.0051]}$	$+0.0094_{[0.0062,0.0126]}$	$+0.0011_{[-0.0022,0.0044]}$
Clean Solution	$-0.0005_{[-0.0033,0.0024]}$	$+0.0093_{[0.0060,0.0128]}$	$-0.0003_{[-0.0034,0.0029]}$
Summary	$-0.0031_{[-0.0063,0.0001]}$	$+0.0089_{[0.0053,0.0125]}$	$+0.0025_{[-0.0010,0.0059]}$
Full Trace	$-0.0123_{[-0.0173,-0.0072]}$	$+0.0131_{[0.0087,0.0176]}$	$+0.0099_{[0.0054,0.0144]}$

Figure 9 separates the correct- and incorrect-response shifts underlying alignment. On average, the teacher assigns lower likelihood than the student to both kinds of response. Correctness alignment measures the difference between those mean shifts.

D.3 Sensitivity to the scored span

Restricting the retained thinking-enabled rows to positions below 1,024 attenuates the alignment trend but leaves correction pressure negative for every view (Figure 10). Thus, the full-profile comparison varies both student states and scoring horizon. This sensitivity check uses already scored tokens rather than generating or rescoring trajectories.

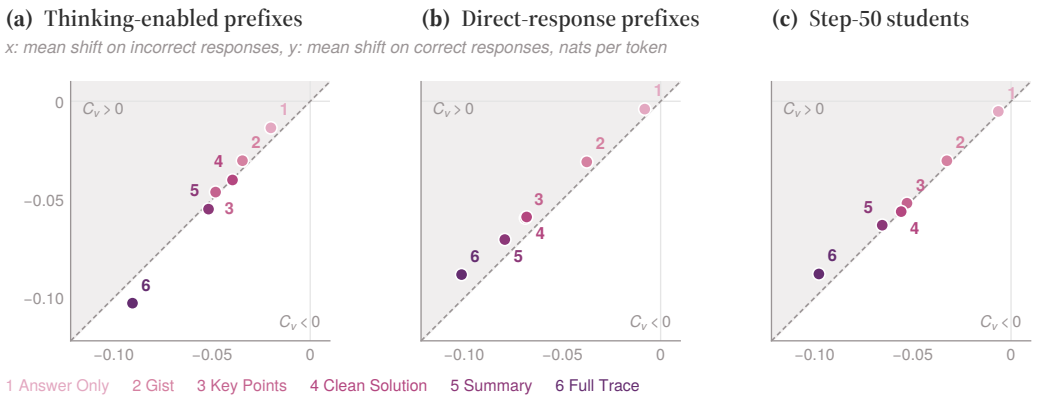


Figure 9: **Correctness alignment separates two negative likelihood shifts.** Each point compares mean per-token log-likelihood shifts on incorrect responses (horizontal axis) and correct responses (vertical axis). The vertical distance from the equal-shift line is C_v. Panels use frozen-base thinking-enabled responses, frozen-base direct responses, and each step-50 student's direct responses, averaging within problems with both outcomes.

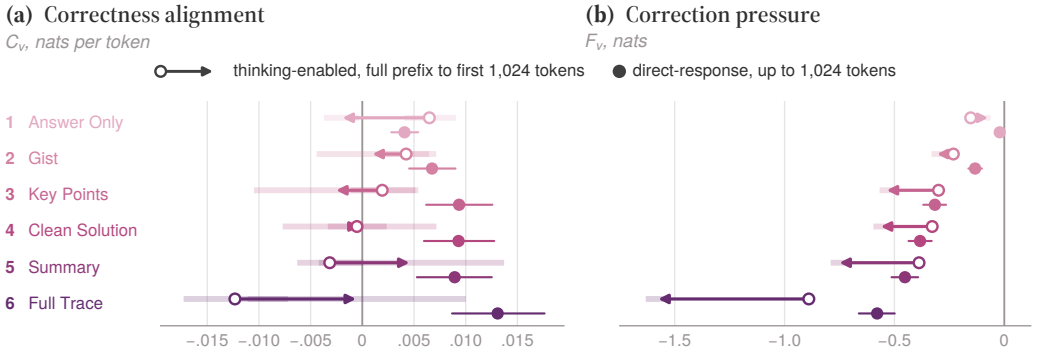


Figure 10: **Profile sensitivity to the scored span.** Arrows connect the full-prefix thinking-enabled profile to the same retained rows below position 1,024. Filled circles show the direct-response profile, capped at 1,024 by design. Light bands and error bars show 95% problem-cluster bootstrap confidence intervals. C_v denotes correctness alignment and F_v the shift at sampled correction markers.

D.4 Interpretation and limits of correctness alignment

Correctness alignment describes how privileged conditioning distinguishes correct from incorrect profiled responses. We give its local reweighting interpretation, then show why the same aggregate diagnostics can accompany different parameter updates.

Reweighting fixed responses. For a mixed-outcome problem x, let μ_x be the empirical distribution assigning equal weight to each profiled completion, $R(y) \in \{0, 1\}$ its correctness indicator, and $h_v(y) = \Delta_v(y)$ its mean shift over the retained span from Section 2.2. Write $p_x = \mathbb{E}_{\mu_x}[R] \in (0, 1)$, so Equation 3 gives $C_v(x) = \mathbb{E}_{\mu_x}[h_v \mid R = 1] - \mathbb{E}_{\mu_x}[h_v \mid R = 0]$. Under the exponential reweighting $\mu_{v,\epsilon}(y) \propto \mu_x(y) \exp(\epsilon h_v(y))$,

$$\frac{d}{d\epsilon} \mathbb{E}_{\mu_{v,\epsilon}}[R] \Big|_{\epsilon=0} = \mathrm{Cov}_{\mu_x}(R, h_v) = p_x(1 - p_x)\, C_v(x).$$

Proof. Differentiating the normalized weighted mean at zero gives $\mathbb{E}[Rh_v] - \mathbb{E}[R]\mathbb{E}[h_v]$. Substituting $\mathbb{E}[Rh_v] = p_x \mathbb{E}[h_v \mid R = 1]$ and expanding $\mathbb{E}[h_v]$ by correctness yields the result. □ Thus, positive $C_v(x)$ means that a sufficiently small positive reweighting raises correctness on that problem's fixed sample. The derivative of mean correctness across problems weights $C_v(x)$ by $p_x(1 - p_x)$,

whereas reported C_v averages problems equally. These two quantities can rank views differently, and reweighting fixed responses is not an OPSD parameter update.

Example: equal alignment and mean KL, opposite transfer. Two equally likely problems have binary one-token answers, with correct answer 1 and student probabilities $p_\theta(1 \mid x_1) = \sigma(\theta)$ and $p_\theta(1 \mid x_2) = \sigma(-2\theta)$, where σ is logistic and $\theta_0 = 0$. Let q_A and q_B be frozen teacher distributions assigning correct-answer probabilities $(0.54, 0.51)$ and $(0.51, 0.54)$, respectively. A same-backbone construction is $p(1 \mid x_i, z) = \sigma(c_i\theta_0 + b_i(z))$, with $c = (1, -2)$, $b_i(\varnothing) = 0$, and $b_i(z_v) = \operatorname{logit} q_v(1 \mid x_i)$. Both contexts have expected teacher accuracy 0.525 and equal mean teacher–student KL because their problem-level distributions are swapped. For a profile containing both answers on each problem, they also have equal alignment at initialization,

$$C_A = C_B = \tfrac{1}{2}\Big(\log \tfrac{0.54}{0.46} + \log \tfrac{0.51}{0.49}\Big) \approx 0.100,$$

since $C_v(x) = \operatorname{logit} q_v(1 \mid x)$. For $\mathcal{L}_v(\theta) = \frac{1}{2}\sum_i D_{\mathrm{KL}}\big(q_v(\cdot \mid x_i) \,\|\, p_\theta(\cdot \mid x_i)\big)$, the gradients $G_v = \mathcal{L}_v'(0)$ are

$$G_A = \tfrac{1}{2}\big[-0.04 + 0.02\big] = -0.010, \qquad G_B = \tfrac{1}{2}\big[-0.01 + 0.08\big] = +0.035,$$

while expected student accuracy $\mathcal{A}(\theta) = \frac{1}{2}[\sigma(\theta) + \sigma(-2\theta)]$ has $\mathcal{A}'(0) = -1/8$. A gradient-descent step of size η changes accuracy by $-0.00125\,\eta + O(\eta^2)$ under A and $+0.004375\,\eta + O(\eta^2)$ under B, giving opposite signs for sufficiently small positive η. Every vocabulary term lies below the 0.05 clipping threshold near initialization, so the first-step conclusion also holds for the implemented clipped divergence. The difference comes from shared parameters: fitting x_1 moves x_2 in the opposite direction with twice the sensitivity, which the aggregate diagnostics do not capture.

E Intervention Evidence

The eleven direct-response control configurations comprise two other-problem references, three trace openings, four correction-marker edits, and two prompt-template changes. We give their construction details below, followed by the loss-support intervention under thinking-enabled training.

E.1 Reference substitutions and trace openings

Other-problem controls replace both the reasoning body and canonical answer with a length-matched view from another problem. Trace openings truncate genuine Full Trace at a token boundary to each problem's own Key Points, Clean Solution, or Summary length (means 298, 523, and 870 Qwen3 tokens across 2,112 problems). All cuts remove the canonical answer section and later reasoning, changing content and length together. None contains a boxed answer.

Main Table 2 gives the thinking-enabled outcomes at step 100. Both other-problem replacements lower accuracy after within-family Holm correction. At development-selected checkpoints, only the Key Points replacement's accuracy difference has an interval below zero. Table 17 adds matched single-seed direct-response comparisons. All three trace openings have positive mode interactions, while the intervals for both other-problem interactions include zero.

E.2 Correction-marker weights and prompt replacements

Marker edits for Full Trace and Clean Solution either exclude marker positions and divide by the retained-token count, or multiply their losses by 0.1 and divide by the original supported-token count. After the 0.05 per-vocabulary-term clip, weighted token losses are averaged within each completion, then equally across the effective batch of 32.

Prompt swaps for Full Trace and Clean Solution replace both student and teacher training wording with the original OPSD templates (Zhao et al., 2026a). Evaluation retains the local student prompt (Appendix D). During training, both versions use one user message and a generation prompt, with student thinking off and teacher thinking on. The replacement templates preserve the original wording and blank lines, with long lines wrapped for display.

Table 17: **Reference-construction effects depend on evaluation mode.** Step-100 Avg@4 differences, in percentage points with paired 95% problem-cluster bootstrap intervals. Every row uses one training seed. Trace openings are compared with seed-0 FULL TRACE, and other-problem references with seed-0 KEY POINTS or CLEAN SOLUTION, respectively. The interaction is the direct-response contrast minus the thinking-enabled contrast on the same problems.

Configuration	Thinking enabled	Direct response	Interaction
Opening, KEY POINTS length	$-0.20_{[-2.02,1.69]}$	$+9.05_{[6.12,12.04]}$	$+9.24_{[5.86,12.63]}$
Opening, CLEAN SOLUTION length	$-0.59_{[-2.28,1.11]}$	$+11.33_{[8.40,14.32]}$	$+11.91_{[8.59,15.30]}$
Opening, SUMMARY length	$-1.11_{[-2.86,0.65]}$	$+7.81_{[5.08,10.55]}$	$+8.92_{[5.66,12.24]}$
Other-problem KEY POINTS	$-2.15_{[-3.97,-0.39]}$	$-3.91_{[-6.64,-1.17]}$	$-1.76_{[-5.01,1.50]}$
Other-problem CLEAN SOLUTION	$-1.95_{[-3.78,-0.20]}$	$-0.39_{[-3.06,2.28]}$	$+1.56_{[-1.76,4.82]}$

Original OPSD prompts used for replacement

Student

```
Problem: {question}

Please reason step by step, and put your final answer within \boxed{}.
```

Teacher

```
Problem: {question}

Here is a reference solution to this problem:
=== Reference Solution Begin ===
{reference text}
=== Reference Solution End ===


After reading the reference solution above, make sure you truly
understand the reasoning behind each step — do not copy or paraphrase
it. Now, using your own words and independent reasoning, derive the
same final answer to the problem above. Think step by step, explore
different approaches, and don't be afraid to backtrack or reconsider if
something doesn't work out:

Please reason step by step, and put your final answer within \boxed{}.
```

E.3 LOSS-SUPPORT INTERVENTION AND CHECKPOINT SENSITIVITY

The one-seed FULL TRACE experiment compares the three loss windows in Section 5. The four 256-token distributed windows are centered near fractions 0.125, 0.375, 0.625, and 0.875 of the reasoning span before `</think>`. All three score highest at step 25 in Table 18, although development selection chooses step 50 for EARLY-1K and step 25 for the alternatives. All paired same-step loss-window contrasts in Table 2 have intervals that include zero.

Table 18: **Loss-support checkpoint sensitivity.** Thinking-enabled FULL TRACE Avg@4 in percent, from one training seed. Paired differences and intervals are in Table 2.

Support	Step 25	Step 50	Step 100
Early-1K	78.26	75.13	72.79
First-4K	79.23	74.28	71.22
Distributed-1K	79.23	75.65	71.48