THE RIGHT DIRECTION AT THE RIGHT STATE: DISSECTING ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) offers a route to improved reasoning by letting a context-augmented copy of the model supervise its own generations. What makes this additional context useful for learning? We investigate this question through controlled studies of mathematical reasoning with Qwen3 thinking models. Solutions to other problems and a single fixed reference can produce useful training effects, while target-specific solutions outperform other-problem references at both evaluated scales. The central finding is that useful guidance depends on the student's reasoning so far. Exchanging reference-induced changes in token preferences between attempts at the same problem reduces continuation accuracy across both scales, with the amount of distributional change matched. On shared Base histories, thinking-only use of the trained policy retains about 90% of its observed full-policy gain. On MATH-500, alignment improves correct completion within 4k tokens by 4.30–4.80 points and reduces mean token use under a 32k cap by 10.7–12.5% relative to exchanged guidance. Together, these findings distinguish the reference a teacher reads from the guidance a student can use. Reference content shapes the learning benefit; the student's evolving reasoning determines the usefulness of the induced guidance. This trajectory-conditioned view connects local supervision during training to successful reasoning completion.

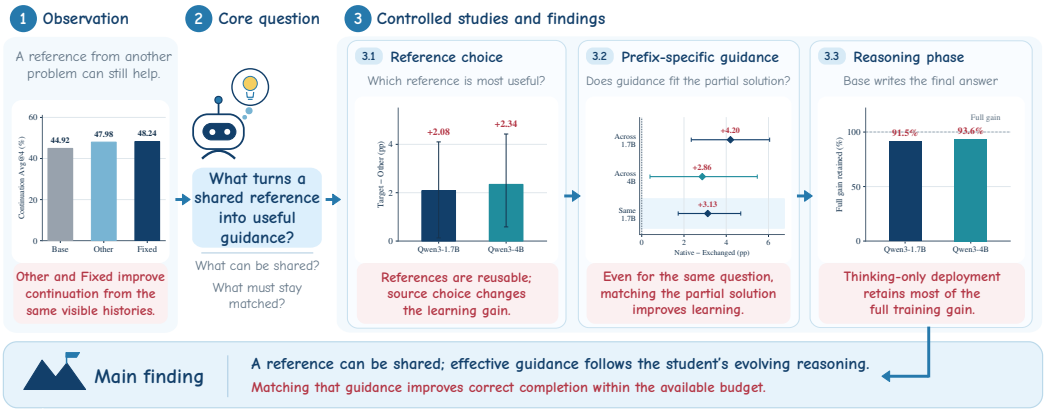

Figure 1: **Reference value, prefix-specific guidance, and reasoning completion.** The initial observation uses Qwen3-1.7B. The study panels show the Target–Other advantage, selected guidance-pairing contrasts, and gains retained by thinking-only deployment from shared Base histories. Their numbers mark the visual sequence; Sections 4–6 develop the corresponding findings.

1 INTRODUCTION

On-policy distillation (OPD) trains language models with teacher feedback on their own generated trajectories (Agarwal et al., 2024; Gu et al., 2024). Its combination of student-generated experience and dense supervision has made it a practical approach to reasoning post-training (Li et al., 2026c; Feng et al., 2026). On-policy *self*-distillation (OPSD) constructs the teacher from the learner itself, supplying additional context such as a reference solution or feedback (Zhao et al., 2026a; Hübotter

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.

et al., 2026). This extends learning with privileged information (Vapnik & Izmailov, 2015; Lopez-Paz et al., 2016): training-time information supervises reasoning that the student must eventually perform unaided.

A central question is how a reference becomes useful supervision. Providing a worked solution changes both the information available to the teacher and the context in which it predicts the student's next token. Recent work shows that references from other problems can also improve self-distillation (Ichihara et al., 2026). At the same time, privileged references can distort deliberation, induce style shifts, and impair long-budget reasoning (Pan et al., 2026; Kaur et al., 2026; Shen et al., 2026a). Analyses of privileged bias and token credit further question when likelihood-based guidance tracks task success (Harne et al., 2026; Nguyen et al., 2026). Together, these observations make the interaction between a reference and the student's ongoing reasoning a natural object of study.

Our central question is whether ***reusing a reference also permits reusing the local feedback it produces***. The same worked example is interpreted anew as the student's partial solution changes. We test the learning value of that correspondence by reassigning only the example-induced change in token preferences, while retaining the receiving prediction and matching the target's departure from it. This distinguishes a reusable teaching input from the prefix-specific supervision through which it shapes learning.

We investigate this distinction in three stages. First, source comparisons and exact-prefix continuation show that reference reuse and source sensitivity coexist: unrelated references can yield useful learning, and Target outperforms Other at both scales. Comparisons with probability sharpening and suppression of unlikely tokens motivate a closer examination of the feedback itself. Second, we exchange only the reference-induced change in token preferences, preserving the receiving state's question-conditioned prediction and matching the change's distributional size. The native assignment improves continuation both across problems and, crucially, between different attempts at the *same* problem. Holding the question fixed reveals this prefix-specific benefit at both scales. Changing the optimization objective from JSD to reverse KL preserves the pairing advantage.

Finally, policy swaps at the end of thinking connect this training-time relationship to generation. Using the trained policy during reasoning and the initial policy for answer emission retains most of the full improvement. Completion measurements show more attempts reaching correct terminal answers within budget. These results support a *trajectory-conditioned* account of self-distillation: a reference supplies feedback whose learning value depends on the reasoning it accompanies. Our contributions are threefold:

- **We separate reference reuse from reference value.** Source-specific continuation reveals useful non-target supervision and a consistent Target–Other advantage. A reference's capacity for reuse coexists with differences in the learning benefit its assignment provides.
- **We establish the learning value of feedback–prefix correspondence.** Preserving native assignments improves continuation when recipient predictions and reference-relative target distances are controlled. Same-question tests give a 2.93–3.13-point advantage at both scales; cross-question gains persist under reverse-KL training.
- **We connect supervisory assignment to reasoning completion.** On shared Base histories, thinking-only deployment retains about 90% of the observed full-policy gain. MATH-500 extends this account to complete problem solving: alignment raises short-budget accuracy and reduces generated tokens, with accuracy differences narrowing at larger budgets.

2 Related Work

On-policy and contextual distillation. Knowledge distillation transfers teacher predictions into a student (Hinton et al., 2015; Kim & Rush, 2016). OPD queries the teacher on student-generated prefixes, aligning supervision with learner-visited states (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024). Instead of requiring another stronger teacher model, contextual self-distillation builds the teacher from the learner with additional information (Zhao et al., 2026a; Hübotter et al., 2026; Snell et al., 2022; Ye et al., 2026). Recently, OP^{2}SD demonstrates that other-problem references can remain useful (Ichihara et al., 2026). We build on that source-level observation by testing how the resulting guidance depends on the student's partial solution.

The reliability of dense supervision. Analyses of thinking models link privileged context to altered deliberation and style (Kaur et al., 2026; Pan et al., 2026). RLSD anchors updates to outcomes (Yang et al., 2026a); Purified OPSD separates a reference-only component (Shen et al., 2026a); reflective and peer-conditioned methods restructure the teacher's context (Zhao et al., 2026b; Yu et al., 2026c). Entropy, predictive overlap, and prefix drift characterize supervision reliability (Li et al., 2026c; Jin et al., 2026; Yang et al., 2026b), while outcome verification supports teacher gating and adaptation (Zhang et al., 2026; Ding et al., 2026). Our intervention tests the functional value of the signal–prefix relation after controlling the reference prediction and contextual divergence.

States, update directions, and computation. Few-query studies examine how rollouts expose teacher supervision (Fu et al., 2026; Hou et al., 2026); teacher-free adaptation studies suppression of unlikely tokens (Ding & Zhang, 2026). Parameter analyses identify sparse or low-dimensional update structure (Yu et al., 2026b; Shen et al., 2026b), and Direct-OPD transfers a pre-/post-RL likelihood shift (Feng et al., 2026). We manipulate context-induced shifts in output space, then use policy crossovers to locate their behavioral consequences. This complements studies of training-state distributions (Nie, 2026) and reasoning-budget allocation (Snell et al., 2025; Muennighoff et al., 2025; Caldarella et al., 2026). Appendix F expands these distinctions.

3 PRELIMINARIES

We introduce the distributions and reference-relative representation used in our study of contextual on-policy self-distillation.

Student and self-teacher. Given a problem x, the student samples a response y, visiting states $s_t = (x, y_{<t})$. On-policy distillation evaluates a teacher on these student-generated prefixes (Agarwal et al., 2024; Gu et al., 2024). In contextual self-distillation, the teacher additionally receives information unavailable to the student (Zhao et al., 2026a; Hübotter et al., 2026). We consider a frozen-teacher setting: the trainable student is $p_\theta(\cdot \mid s_t)$, while a frozen copy of its initialization defines both the unaugmented reference $q_0(\cdot \mid s_t)$ and the context-conditioned teacher $q_C(\cdot \mid s_t)$. Both teacher views evaluate the same student-generated prefix, with C supplied only to the contextual view. Although its weights remain fixed, the teacher distribution depends on the question and evolving prefix.

Distillation objective. For distributions p and q over a shared vocabulary $\mathcal{V}$, our main experiments use symmetric Jensen–Shannon divergence, a bounded, symmetric measure that also supports reference-relative target calibration:

$$J(p, q) = \tfrac{1}{2} D_{\mathrm{KL}}(p\|\mu) + \tfrac{1}{2} D_{\mathrm{KL}}(q\|\mu), \qquad \mu = \tfrac{1}{2}(p + q). \tag{1}$$

For sampled response positions $\mathcal{B} = \{(i, t)\}$, with trajectory index i and token index t, we minimize a token-level divergence $\mathcal{D}$ to target q:

$$\mathcal{L}_{\mathcal{B}}(\theta; q) = \frac{\sum_{(i,t)\in\mathcal{B}} w_{i,t}\, \mathcal{D}(p_\theta(\cdot \mid s_{i,t}), \mathrm{sg}[q(\cdot \mid s_{i,t})])}{\sum_{(i,t)\in\mathcal{B}} w_{i,t}}. \tag{2}$$

The main studies set $\mathcal{D} = J$; the objective comparison sets $\mathcal{D}(p, q) = D_{\mathrm{KL}}(p\|q)$. Reference-relative target matching uses J in both cases. Here $w_{i,t} \in \{0, 1\}$ selects eligible response positions and sg denotes stop-gradient. The loss uses the full vocabulary. During differentiation, sampled tokens, masks, and teacher targets are held fixed; the student next-token distribution receives gradients.

Reference-relative teacher directions. To describe the contextual change separately from the unaugmented prediction, for tokens $a, v \in \mathcal{V}$ and natural logarithms, we define

$$\begin{aligned} r_C(s, a) &= \log q_C(a \mid s) - \log q_0(a \mid s), \\ d_C(s, a) &= r_C(s, a) - \mathbb{E}_{v\sim q_0(\cdot|s)} r_C(s, v). \end{aligned} \tag{3}$$

Centering removes an additive logit offset without changing the induced distribution:

$$q_C(a \mid s) \propto q_0(a \mid s) \exp d_C(s, a). \tag{4}$$

We refer to $d_C(s)$ as the *teacher direction*: a centered log-probability displacement in the shared output vocabulary. Equation 4 separates the contextual reweighting from the state's reference prediction. A fixed context can induce different directions at different states; this representation lets us study their pairing while retaining each recipient's own reference distribution.

4 What Does the Reference Contribute?

4.1 Experimental setup

Models, data, and training. We study Qwen3-1.7B and Qwen3-4B with thinking enabled (Yang et al., 2025). Each condition starts from its model's initialization, uses three training seeds, and runs 100 LoRA steps with rank 64, scaling 128, learning rate 2×10^{-6}, and a 4,096-token training completion cap. Sources and shaping controls use 101 screened NuminaMath-1.5 problems. Cross-problem pairing uses 98 of these in 49 fixed within-family pairs; same-question pairing uses the same 98 problems, each with two separately sampled attempts. Both scales undergo same-question pairing; the smaller model also tests reverse-KL optimization. Each training arm samples its own evolving trajectories. Appendix A specifies the frozen inputs, sampling, and implementation.

Two evaluation channels. Complete-response evaluation uses 30 AIME25 and 30 HMMT25 questions at 4k, and 500 MATH-500 questions at up to 32k (Section 6.3). For mechanism analysis, we freeze Base-generated histories and let every evaluated policy continue exactly the same token IDs. Each policy re-encodes that history with its own weights. Histories contain at most 4k response tokens and receive up to 4k additional tokens; already terminated histories retain their outcomes. Thus, the experiment fixes visible reasoning history while varying the continuation policy. Source/shaping evaluation uses the same 128 DAPO-derived problems at both scales, with model-specific Base histories. Pairing uses non-overlapping cohorts of 256 and 128 problems, each with four histories per problem. Same-question tests reuse these model-specific histories; reverse KL reuses the 256-problem cohort. Public shared-prefix evaluation reuses the 60 contest questions. Cohorts are selected deterministically before inspecting their model outputs. Table 2 records all experiment reuse.

Outcomes and uncertainty. A response is correct when it exposes a verifiable final answer after thinking closes, within budget. For N problems, K histories $h_{i,k}$ per problem, and training seeds $r \in \mathcal{R}$, let $R(g_r; h_{i,k}) \in \{0, 1\}$ denote correctness under seed-r policy regime g_r. We average:

$$\widehat{A}_g = \frac{1}{|\mathcal{R}|N} \sum_{r \in \mathcal{R}} \sum_{i=1}^{N} \left[\frac{1}{K} \sum_{k=1}^{K} R(g_r; h_{i,k}) \right]. \tag{5}$$

We report Avg@K as per-sample correctness, alongside valid completion, wrong-valid completion, and budget censoring. Endpoint comparisons use a common 20,000-resample paired problem/training-seed bootstrap, with 95% intervals. The temporal curves report point estimates at five recorded continuation horizons; Appendix D.3 retains their original uncertainty summaries. Histories are averaged within problem before resampling. Appendix E.4 gives the paired procedure and finite-seed interpretation; Appendix E.8 records the analysis-specific interval conventions.

4.2 References can be reused, while source choice changes the gain

We first vary what the teacher reads. Target supplies the current problem and its solution; Other supplies another problem–solution pair through a fixed derangement; Fixed repeats one proof that 30 divides $n^5 - n$. The wrapper, student input, initialization, data, and training recipe are shared. The source-specific experiments extend the reference-reuse observation of Ichihara et al. (2026) to explicit thinking and controlled continuation.

Reference reuse and a target-specific advantage coexist. Figure 2(a) separates the reference sources on shared histories: Other and Fixed raise 1.7B accuracy by 3.06 and 3.32 points, while Target gains 5.14. Target is also the strongest source in the 4B comparison, gaining 2.54 points versus 0.20 for Other and 0.85 for Fixed (absolute rates in Appendix B.1, Table 5). The direct Target–Other contrast is positive at both scales: 2.08 and 2.34 points (Figure 2(b)), with the same

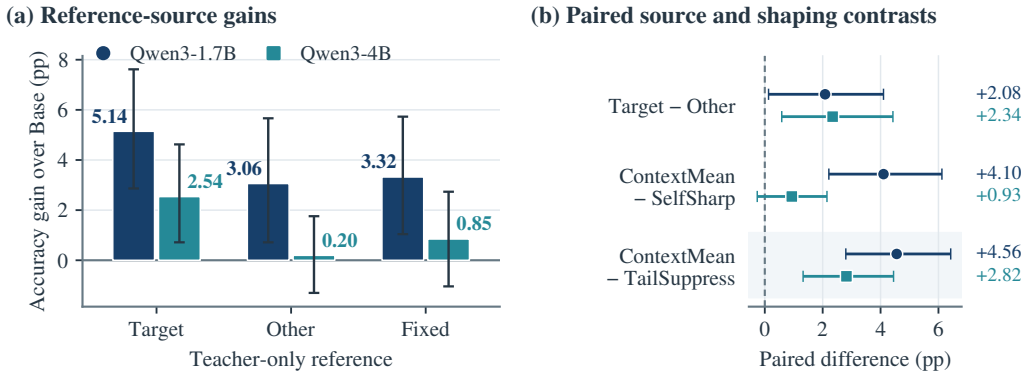

Figure 2: **A reference can be reused while its source changes the learning benefit.** (a) Source-specific accuracy gains over Base from identical visible histories. (b) Direct paired contrasts distinguish the Target–Other increment from the two shaping comparisons. Both models use the common 128-problem source/shaping cohort, with model-specific Base histories. Bars and filled markers show means; whiskers are paired 95% intervals. ContextMean averages the three source policies' evaluation outcomes.

ordering in every training seed. Non-target gains are more pronounced in the smaller model. These results separate two properties of a reference: it can be reusable across questions, and its identity can still change how much the student learns.

The complete-response source study gives positive point estimates for all three references on the contest questions, with larger sampling uncertainty under the 4k cap. Parameter recomposition provides a complementary test of the functional value of non-target-trained updates, while budget sweeps examine their performance over additional inference compute (Appendices B.3–B.4). The shared-prefix comparison sharpens the present question by attributing the source differences to policies acting from the same visible histories.

4.3 Contextual gains extend beyond tail suppression

A worked example may reinforce common tokens or suppress unlikely continuations. We compare its trained effect with two explicit rules. SelfSharp sharpens the unaugmented frozen target:

$$q_{\text{sharp}}(a \mid s) \propto q_0(a \mid s)^{1/\tau}, \quad 0 < \tau < 1. \tag{6}$$

The temperature is calibrated before training to the mean initial contextual JSD on 16 training problems, then held fixed. TailSuppress assigns entropy-adaptive negative advantages to the lowest-probability 20% of sampled tokens within each response, following Ding & Zhang (2026). Appendix E.3 specifies both rules. Here ContextMean is the equal-weight average of the three source policies' evaluation outcomes.

ContextMean exceeds tail suppression by 4.56 and 2.82 points across the two scales. Its advantage over sharpening is 4.10 points in the smaller model and 0.93 in the larger, where their estimates are closer (Figure 2(b)). The paired intervals display the precision of each contrast directly. Thus the clearest repeated shaping contrast is with the tail-token rule, while the full source and sharpening results reveal how the size of contextual gains varies. This motivates a more direct question: *which property of the contextual change makes it useful at the student's current state?*

5 Useful Guidance Follows the Student's Reasoning

We fix the reference and vary the assignment of its induced feedback. The same proof is shared across all problems, holding its content and style constant while the teacher interprets different question–prefix states. Cross- and same-question exchanges test how the feedback depends on those states; the objective comparison tests how this assignment affects learning under a different loss.

5.1 EXCHANGING GUIDANCE WHILE PRESERVING THE RECEIVING PREDICTION

Aligned uses the contextual distribution at the receiving state: $q_A(\cdot \mid s_{i,t}) = q_C(\cdot \mid s_{i,t})$. StatePermuted takes a donor direction from $s_{j,t}$, centers it under the recipient reference, and constructs

$$\widetilde{d}_{j\to i,t}(a) = d_C(s_{j,t}, a) - \mathbb{E}_{v\sim q_0(\cdot|s_{i,t})} d_C(s_{j,t}, v), \tag{7}$$

$$q_P(a \mid s_{i,t}) \propto q_0(a \mid s_{i,t}) \exp\Big(\alpha_{i,t}\widetilde{d}_{j\to i,t}(a)\Big). \tag{8}$$

Here i and j index recipient and donor trajectories. Each recipient retains q_0; its scale $\alpha_{i,t} \geq 0$ is solved against the actual full-vocabulary divergence, with all arguments conditioned on $s_{i,t}$:

$$|J(q_0, q_P) - J(q_0, q_A)| \leq 0.05 J(q_0, q_A) + 10^{-7}. \tag{9}$$

Matching controls each target's divergence from the receiving frozen reference. Both candidate targets share accepted positions within a sampled batch; each training arm follows its own on-policy trajectories. Student–target distances and gradients express the optimization consequences of the assignment. Appendix E reports numerical and coverage diagnostics.

For the cross-problem test, donor and recipient come from the same construction family, at the same response index and generation phase. This matches coarse task structure and phase while varying the joint question–prefix state. **Preserving the native assignment yields higher continuation accuracy at both scales.** The connected endpoints in Figure 3(a) compare each policy with its own Base. Aligned gains 3.06 points in each cohort; its direct advantage over StatePermuted is 4.20 and 2.86 points (Figure 3(b); Table 1). All six paired seed differences favor Aligned. The public contest histories show the same aggregate ordering, with gaps of 0.69 and 2.08 points; Appendix B.2 reports their wider intervals and separate benchmark evaluation protocol.

5.2 DIFFERENT ATTEMPTS AT THE SAME QUESTION NEED DIFFERENT GUIDANCE

A student state includes both the question and its partial solution. To hold the question fixed, we sample two independent attempts at each of the same 98 training problems, forming states $s_t^{(a)} = (x, y_{<t}^{(a)})$ and $s_t^{(b)} = (x, y_{<t}^{(b)})$. The teacher sees the same question and fixed reference for both. After the prefixes diverge, IntraPermuted exchanges their directions at equal response indices and phases. IntraAligned retains the original assignment. Both arms are trained afresh with the same paired-rollout organization, prefix-divergence mask, and matching rule, for three seeds each.

Even within the same problem, feedback is more useful at the prefix that elicited it. The pairing advantage is 3.13 and 2.93 points at the two scales, accompanied by gains over Base of 2.64 and 2.80 points. All six paired training-seed comparisons favor IntraAligned (Figure 3). With question and reference held fixed, this replicated effect identifies the generated prefix as a functional component of useful supervision. Appendix C.2 reports the absolute outcomes and seed-level comparisons.

5.3 THE PAIRING ADVANTAGE PERSISTS UNDER REVERSE-KL OPTIMIZATION

We next vary how the student learns from the constructed targets. Using the cross-question design, we replace the JSD training loss with full-vocabulary $D_{\mathrm{KL}}(p_\theta \| q)$ and train six new policies. The fixed reference, question pairs, phase rule, and recipient-reference JSD matching are retained. **Native pairing remains beneficial under reverse KL.** Its advantage over exchange is 4.04 points, alongside a 3.16-point improvement over Base. The corresponding JSD pairing gap is 4.20 points on the same histories (Table 1). Thus, preserving the guidance–state correspondence improves learning under both tested objectives. Appendix C.3 gives the loss, outcomes, and matching diagnostics.

Together, these interventions distinguish *reference reuse* from *feedback reassignment*. One fixed example yields more useful learning when its induced preferences remain attached to the prefix that elicited them. The cross-question comparison preserves this advantage under both tested objectives. We next ask *where this learned advantage acts during generation.*

Table 1: **Native guidance is more useful at its original prefix.** Paired Avg@4 effects in percentage points with 95% intervals. Each row compares separately trained native and exchanged arms. The 1.7B and 4B evaluations use 256 and 128 problems, respectively; same-question and loss comparisons reuse the corresponding histories. Target-strength matching uses JSD throughout.

Exchange	Model	Loss	Native − Base	Native − Exch.	Exch. − Base
Across questions	1.7B	JSD	+3.06 [+1.30, +4.95]	+4.20 [+2.34, +6.05]	-1.14 [−2.80, +0.59]
Across questions	4B	JSD	+3.06 [+0.59, +5.60]	+2.86 [+0.39, +5.47]	+0.20 [−2.67, +2.93]
Same question	1.7B	JSD	+2.64 [+1.07, +4.30]	+3.13 [+1.73, +4.69]	-0.49 [−2.12, +1.11]
Same question	4B	JSD	+2.80 [+0.20, +5.60]	+2.93 [+0.65, +5.14]	-0.13 [−2.99, +2.86]
Across questions	1.7B	Reverse KL	+3.16 [+1.43, +4.98]	+4.04 [+2.08, +6.05]	-0.88 [−3.06, +1.17]

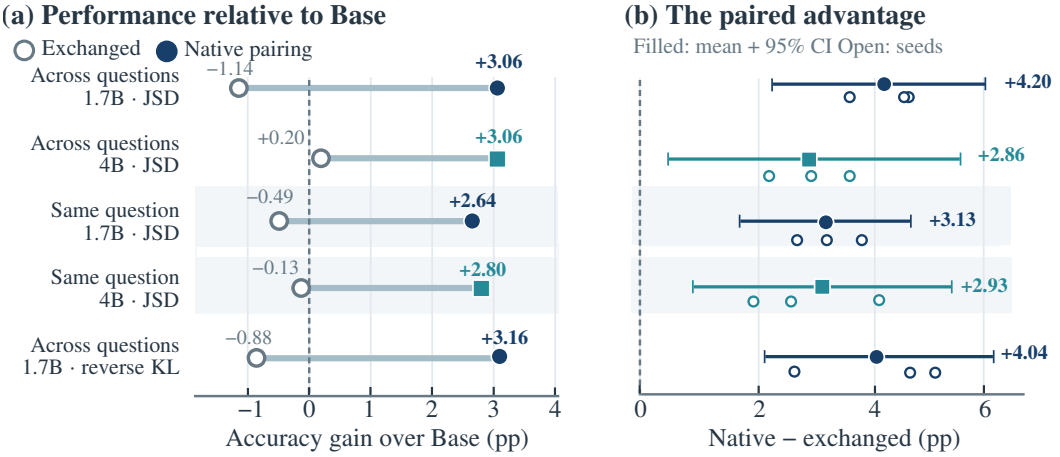

Figure 3: **Guidance is more useful at the partial solution that elicited it.** (a) Native and exchanged-direction policies relative to their corresponding Base, shown as connected point estimates. (b) The directly estimated native–exchanged difference with paired 95% intervals and individual training-seed effects (open markers). Shaded same-question rows fix problem identity; the RKL row changes the training loss. Target-strength matching uses recipient-reference JSD in every row. Each comparison uses newly trained matched arms and the indicated model's fixed cohort.

6 How Matched Guidance Improves Reasoning Completion

6.1 The trained reasoning policy retains most of the positive gain

From shared Base histories, we separate the reasoning policy from the policy that writes the exposed answer. A regime uv uses u until the actual `</think>` boundary and v afterwards; the closing token belongs to u. Every regime re-encodes the exact prefix at the switch and uses the remaining budget. Terminated histories remain absorbed. Let V_{uv} denote continuation accuracy, A the original cross-problem Aligned policy, P StatePermuted, and B Base.

Locating the positive training effect. The A/B matrix directly tests how much of Aligned's improvement survives when the final answer is written by Base. **Using Aligned only during thinking retains 91.5% and 93.6% of the full gain.** Thinking-only deployment (AB) gains 2.80 and 2.86 points over BB, compared with 3.06 points for full AA. Answer-only deployment (BA) gains 0.55 and 0.46 points (Figure 4(a)). The retention percentages are the ratios $(V_{AB} - V_{BB})/(V_{AA} - V_{BB})$ of these point estimates. They quantify the improvement retained by a deployment regime.

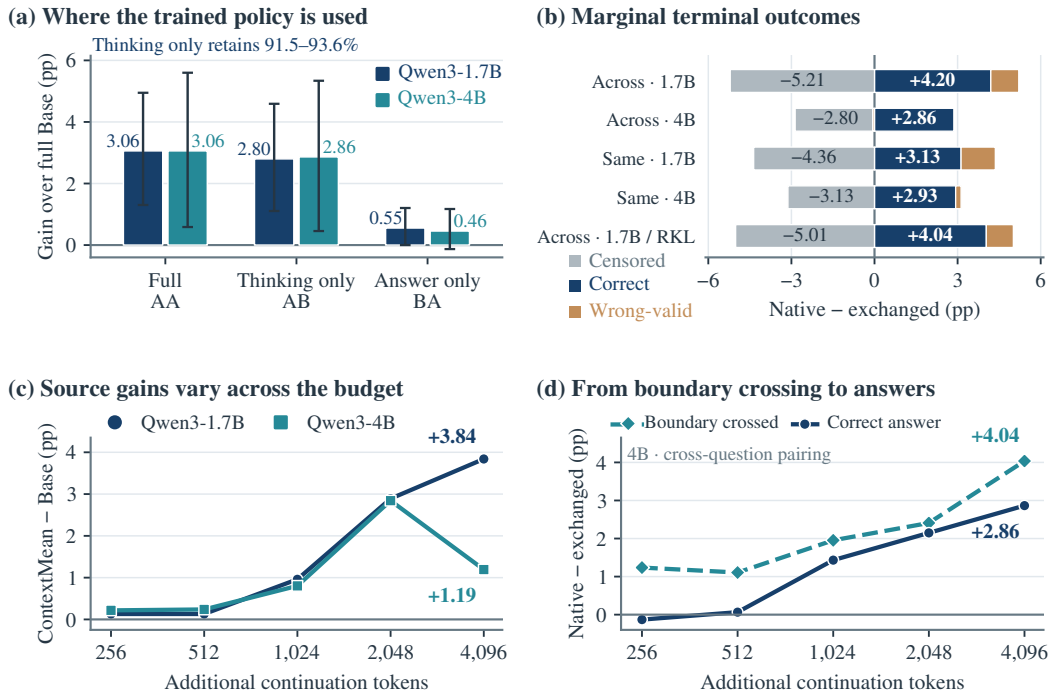

Figure 4: **Matched guidance improves reasoning-stage execution and timely completion.** (a) Gains over full Base from full (AA), thinking-only (AB), and answer-only (BA) Aligned deployment; whiskers are paired 95% intervals. (b) Native–exchanged changes in marginal terminal frequencies for both pairing scopes and the RKL comparison, partitioned into correct, wrong-valid, and censored outcomes. (c) ContextMean–Base accuracy gaps on the common source/shaping cohort. (d) Boundary-crossing and correct-answer gaps on the separate 4B pairing cohort. Curves show point estimates at the five recorded continuation horizons, joined on a log-base-2 token axis.

Locating the pairing effect. The complementary A/P matrix places the pairing advantage in the same phase. Its total difference admits two conditional decompositions:

$$\begin{aligned} V_{AA} - V_{PP} &= (V_{AA} - V_{PA}) + (V_{PA} - V_{PP}) \\ &= (V_{AP} - V_{PP}) + (V_{AA} - V_{AP}). \end{aligned} \tag{10}$$

Changing the thinking policy contributes 3.74–3.78 and 2.54–2.80 points across the two scales; changing the answer policy contributes 0.42–0.46 and 0.07–0.33. Each range spans the two conditional point estimates. Both the positive training effect and the native-pairing advantage are thus expressed principally through the policy used during reasoning. That phase determines the content and timing of the prefix delivered to the answer policy, including whether the boundary is reached. Appendix D.1 reports all conditional effects and the analysis restricted to actual answer-policy calls.

6.2 Matched guidance increases correct completion

We next examine the terminal outcomes of the same initial histories. Across the five pairing comparisons, native guidance increases correct answers by 2.86–4.20 points and valid completion by 2.80–5.21 points, with corresponding declines in censoring. Figure 4(b) separates these changes into correct, wrong-valid, and censored outcomes. In particular, same-question pairing adds 3.13 and 2.93 points in correctness across the two scales, while reverse-KL pairing adds 4.04. The accompanying wrong-valid differences are 1.24, 0.20, and 0.98 points. Preserving the guidance–prefix relation thus improves correct completion from the same histories. Table 15 in Appendix D.2 provides the numerical breakdown.

The benefit unfolds over the continuation budget. In the source/shaping cohort, the 1.7B ContextMean–Base gap grows from 0.13 points at 256 new tokens to 3.84 at 4,096. The 4B gap

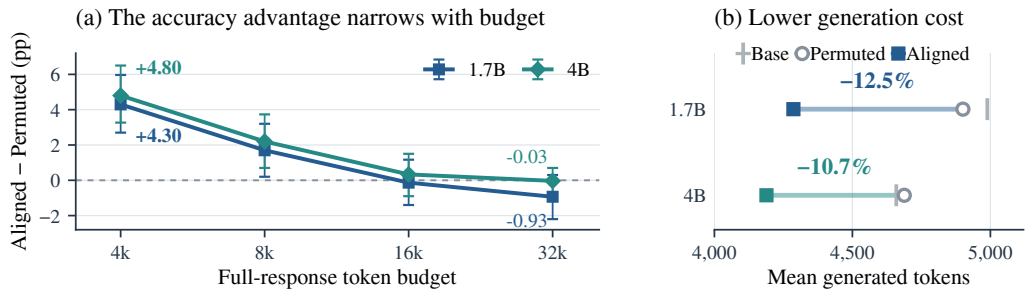

Figure 5: **Matched guidance improves short-budget accuracy and lowers token use.** MATH-500, all 500 questions, two samples per policy. (a) Avg@2 Aligned–Permuted gaps across four full-response budgets; whiskers are pointwise paired 95% intervals. (b) Mean generation length under the 32k cap; reductions are relative to Permuted.

reaches 2.84 points at 2,048 and narrows to 1.19 at 4,096 (Figure 4(c)). The separate 4B pairing cohort shows a 1.24-point boundary-crossing gap at 256 tokens, when the correct-answer gap is near zero; by 4,096, the two gaps reach 4.04 and 2.86 points (Figure 4(d)). Earlier boundary separation accompanies later accumulation of correct answers. Across the same-question and reverse-KL comparisons, native guidance combines higher correctness with 86.7–96.8 fewer continuation tokens. Appendix D.3 gives all cohort-specific windows, token allocations, and interval versions.

Together, the crossover and trajectory measurements connect prefix-specific training supervision to correct completion of an ongoing solution. We next test how this relationship appears when the policy generates its entire solution and receives a larger inference budget.

6.3 Matched guidance improves full-solution efficiency

We evaluate the same cross-question JSD checkpoints on all 500 MATH-500 questions (Hendrycks et al., 2021), generating from each question alone with a 32,768-token cap. Correct completion by budget requires natural termination, closed thinking, and a correct exposed final answer. The prespecified one-sample 32k Aligned–Permuted differences are -0.93 and -0.07 points across the two scales. Two complete rounds supply Avg@2, including the first round; all four prespecified budgets use the same responses.

Matched guidance improves the use of limited inference computation. At 4k, Aligned exceeds Permuted by 4.30 and 4.80 points and Base by 4.83 and 5.50; at 8k, the pairing advantages are 1.70 and 2.20 points. They reach -0.93 and -0.03 at 32k (Figure 5(a)). A supplementary paired analysis quantifies this budget dependence: the 4k advantage exceeds the 32k difference by 5.23 and 4.83 points (Table 29). Under the 32k cap, Aligned generates 614 and 499 fewer tokens on average than Permuted, reductions of 12.5% and 10.7% (Figure 5(b)). Together with the crossovers, these results connect prefix-matched training supervision to correct completion within limited inference compute. Only three of 14,000 responses finish their final correct box by 4k yet stop later. Appendices D.4–D.6 report all comparisons and scoring sensitivity.

7 Conclusion

Contextual OPSD learns from the relationship between a reference and the learner's ongoing reasoning. Our controls identify reusable but unequally valuable references, prefix-specific guidance at both scales, and a pairing advantage under JSD and reverse KL. Stage swaps connect these gains to the trained reasoning policy. MATH-500 links this correspondence to higher short-budget correct completion and lower token use, with 32k accuracy differences near zero or slightly negative.

Together, these findings motivate a *trajectory-conditioned* view of supervision: each reference defines local guidance over partial solutions, and preserving this correspondence helps translate limited computation into correct answers. This perspective suggests preserving prefix-specific feedback when compressing references or caching teacher signals. ***A reference can be shared; effective supervision follows the reasoning it accompanies.***

Reproducibility Statement

The distillation formulation is defined in Section 3, and the experimental setup and outcome measure are given in Section 4.1. Target construction and state-pairing interventions are specified in Section 5. Appendix A provides datasets, prompts, training configurations, and evaluation protocols. Appendices B–D report the expanded source, pairing, completion, and MATH-500 results. Appendix E gives the target derivations, realized numerical checks, and paired statistical procedures; Appendix A.6 specifies the computational controls.

AI Use Statement

We used generative AI tools for manuscript polishing and language editing, and for literature retrieval and discovery, including identifying potentially relevant work. The authors take responsibility for the final content of this work, including its text, claims, and references.

Ethics Statement

The study analyzes language-model post-training on mathematical problems and does not report human-subject experiments. Third-party datasets retain their original attribution and terms.

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 *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025. URL https://arxiv.org/abs/2503.04697.

Simone Caldarella, Davide Talon, Rahaf Aljundi, Elisa Ricci, and Massimiliano Mancini. Thinking past the answer: Evaluating harmful overthinking in large reasoning models. arXiv preprint arXiv:2606.02835, 2026.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. arXiv preprint arXiv:2608.01953, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Yi Ding and Ruqi Zhang. Does on-policy distillation really distill? from noisy teacher to self-improvement. arXiv preprint arXiv:2608.31046, 2026.

Zewen Ding, Zezhong Wu, Zhou Tao, Shida Wang, Shizhuo Hou, YongXiang Hua, Haoyu Cao, and Linli Xu. VISTA: Verifier-informed student-to-teacher adaptation for on-policy self-distillation. arXiv preprint arXiv:2608.28306, 2026.

Shiyuan Feng, Huan-ang Gao, Haohan Chi, Hanlin Wu, Zhilong Zhang, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Weak-to-strong generalization via direct on-policy distillation. arXiv preprint arXiv:2607.05394, 2026.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking on-policy distillation of large language models II: One training example. arXiv preprint arXiv:2609.04172, 2026.

Naibin Gu, Qingyi Si, Chenxu Yang, Chuanyu Qin, Junhao Zhou, Peng Fu, Zheng Lin, and Weiping Wang. CompassOPD: Cross-family on-policy distillation via within-family likelihood shifts. arXiv preprint arXiv:2609.10154, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How PI-conditioned teachers break self-distillation. arXiv preprint arXiv:2608.04794, 2026.

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

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation for multilingual math reasoning. arXiv preprint arXiv:2608.05802, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Zhinan Hou, Jiaqi Zhang, Xunliang Cai, and Keyou You. What matters in on-policy distillation? a perspective on data efficiency and data selection. arXiv preprint arXiv:2609.05198, 2026.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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://arxiv.org/abs/2106.09685.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. arXiv preprint arXiv:2601.20802, 2026.

Jaehui Hwang, Sangdoo Yun, Byeongho Heo, and Dongyoon Han. Revisiting complete reasoning traces for post-training. arXiv preprint arXiv:2609.07103, 2026.

Yuki Ichihara, Naoto Iwase, Mohammad Atif Quamar, and Junpei Komiyama. Privileged solutions or context-induced teacher behavior? dissecting on-policy self-distillation. arXiv preprint arXiv:2608.09228, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. arXiv preprint arXiv:2603.07079, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. arXiv preprint arXiv:2607.05184, 2026.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24872–24895. PMLR, 2024. URL https://proceedings.mlr.press/v235/ko24c.html.

Boyan Li, Bingsen Chen, Chenghao Yang, Ping Nie, Chen Zhao, and Xi Ye. Sequential beats joint: On the interplay between on-policy distillation and RLVR. arXiv preprint arXiv:2609.04108, 2026a.

Gengsheng Li, Mao Zheng, Mingyang Song, Jie Sun, Zeyuan Liu, Ruiqi Liu, Qiyong Zhong, Haiyun Guo, Junfeng Fang, and Jinqiao Wang. Data-free on-policy distillation. arXiv preprint arXiv:2609.14193, 2026b.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang 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, 2026c.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision. arXiv preprint arXiv:2608.06296, 2026d.

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 *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Tao Liu, Taiqiang Wu, Mao Zheng, Xuan Luo, Runming Yang, Xuewei Yang, Junjie Wang, and Yujiu Yang. Simple-OPD: Demystifying warm-up for on-policy distillation. arXiv preprint arXiv:2608.06802, 2026.

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.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151/.

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

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. arXiv preprint arXiv:2608.09263, 2026.

Dong Nie. Post-training is about states, not tokens: A state distribution view of SFT, RL, and on-policy distillation. arXiv preprint arXiv:2605.22731, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation. arXiv preprint arXiv:2606.11709, 2026.

Youngrok Park, Sangmin Bae, Hojung Jung, Jongwoo Ko, Yunseon Choi, Young Jin Kim, Pashmina Cameron, Aaron Courville, and Se-Young Yun. Eliciting weak-to-strong generalization with on-policy reverse distillation. arXiv preprint arXiv:2609.08798, 2026.

Rongcan Pei, Zhepei Wei, Shuyao Xu, Xinyu Zhu, Wei-Lin Chen, and Yu Meng. Negative self-distillation: Learning to reason by avoiding flaws. arXiv preprint arXiv:2609.11699, 2026.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of AISTATS*, volume 15, pp. 627–635. PMLR, 2011.

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

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. arXiv preprint arXiv:2607.02234, 2026a.

Zhennan Shen, Yanshu Li, Qingyu Yin, Chak Tou Leong, Zhilin Wang, Yanxu Chen, Rongduo Han, Sunbowen Lee, and Yi R. Fung. On the geometry of on-policy distillation. arXiv preprint arXiv:2606.07082, 2026b.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. arXiv preprint arXiv:2209.15189, 2022.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL https://www.jmlr.org/papers/v16/vapnik15b.html.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. arXiv preprint arXiv:2607.13399, 2026a.

Xiaodong Wang and Peixi Peng. Aha-flow distillation: Flow markers matter in LLM reasoning. arXiv preprint arXiv:2609.07036, 2026.

Xinye Wang, Junxiao Liu, and Shujian Huang. RP-OPSD: Reasoning-pivot-guided on-policy self-distillation for multilingual reasoning transfer. arXiv preprint arXiv:2608.06347, 2026b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Tianle Xia, Lingxiang Hu, Yiding Sun, Linfang Shang, Ming Xu, Lan Xu, Ning Zheng, Wei Xu, and Jie Jiang. Distillation as probability transport: Routed on-policy distillation. arXiv preprint arXiv:2609.08337, 2026.

Han Xiao, Yifan Niu, Dongyi Liu, Chang Luo, and Jia Li. TV-regulated OPD: Direction matters in on-policy distillation. arXiv preprint arXiv:2609.08341, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. arXiv preprint arXiv:2604.03128, 2026a.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-OPD: Efficient and reliable on-policy distillation for long-horizon reasoning. arXiv preprint arXiv:2605.07804, 2026b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. arXiv preprint arXiv:2602.12275, 2026.

Fangxu Yu, Weijia Xu, Michael Xu, Tianyi Zhou, and Zinan Lin. Weak-to-strong on-policy distillation. arXiv preprint arXiv:2607.26246, 2026a.

Guo Yu, Wenlin Liu, Yulan Hu, Hao-Xuan Ma, Jun-Peng Jiang, and Han-Jia Ye. Dense supervision, sparse updates: On the sparsity and geometry of on-policy distillation. arXiv preprint arXiv:2606.13657, 2026b.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures. arXiv preprint arXiv:2605.12652, 2026c.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation. arXiv preprint arXiv:2609.02998, 2026.

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 *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2601.18734.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. ROSD: Reflective on-policy self-distillation for language model reasoning across domains. arXiv preprint arXiv:2605.28014, 2026b.

A Experimental Design and Evaluation Protocols

Guide to the supplementary material. Appendices B–D follow the three empirical questions in Sections 4–6: reference value, state-specific guidance, and reasoning completion. Appendix E supplies the target derivations, realized numerical checks, and statistical procedures; Appendix F relates the interventions to existing mechanisms. This section specifies the shared experimental design.

A.1 Datasets, pairing units, and evaluation cohorts

The source-training set contains 101 accepted examples from the `olympiads_ref` portion of NuminaMath-1.5. Cross-question pairing retains 98 of these examples. Within each construction-family label, examples are ordered deterministically and paired; one example is omitted from each odd-sized family, yielding 49 pairs. These assignments are fixed before rollout success or teacher disagreement is evaluated. The *Other* source uses a separate one-to-one derangement across different construction-family labels and recorded answers. Same-question pairing duplicates each of the same 98 retained problems into two independently sampled response slots, with question and Fixed context held constant. Its 98 pairs permit exchanges after the generated prefixes diverge. Thus, source assignment, cross-question pairing, and same-question pairing manipulate different relations within the same training problem set.

The natural-mathematics evaluation pool contains 4,096 screened English examples from `open-r1/DAPO-Math-17k-Processed`. The pinned revisions are `31dd309567e3da778038cc87d868b6097a3ccf68` for processed DAPO and `1b05109f9e5c1ad06c0663519502416c30b300f8` for NuminaMath-1.5. Evaluation cohorts are selected using fixed namespace-hash orderings after excluding consumed or reserved problems. Normalized exact and number-redacted fingerprints implement exclusion and within-cohort deduplication. These checks address identifiable overlap in the selected data; pretraining exposure and all semantic-template overlap are outside their coverage.

The 1.7B recomposition, source–budget, source/shaping, and pairing cohorts contain 128, 64, 128, and 256 questions. A separate 128-question cohort evaluates 4B pairing, excluding the other consumed and reserved cohorts under the same fingerprint rules. The two scales share the 128 source/shaping question IDs and generate their own Base histories. Same-question and Aligned/Base crossover tests reuse each scale's pairing histories; reverse-KL uses the 1.7B pairing histories. The 64-question source–budget cohort also supports the two-history continuation analysis. Reuse is made explicit in Table 2, so related comparisons remain paired without counting the same questions as independent replications.

Public source transfer uses 30 AIME25 and 30 HMMT25 questions to evaluate frozen source checkpoints. The public pairing test continues fixed histories on these same contest questions. Cohort selection and evaluation rules are fixed before examining task-model outputs for each comparison; the studies are carried out sequentially. Every reported checkpoint is evaluated without test-driven checkpoint selection.

A.2 Data screening and mathematical review

Retained problems pass structural checks for direct target leakage, extraneous meta-instructions, unavailable-media dependence, and invalid answer sentinels. Two independent human reviewers assess problem completeness, reference consistency, verifiability, and possible cross-split duplication; a separate adjudicator examines disagreements. Each initial judgment is produced before the other assessment is shown.

All 101 context-training examples are retained after this process. For the DAPO parent pool, a coverage-oriented 64-row audit accepts 61 examples and rejects three with distinct failure mechanisms. Replacements follow deterministic reserve ordering before task-model outputs are inspected. The remaining parent-pool examples, and evaluation cohorts drawn from them, receive source-level structural screening. The audit describes the inspected cases rather than a random-sample estimate of population label quality. Its scope is therefore distinct from exhaustive question-by-question mathematical review of every evaluation cohort.

Table 2: Experimental coverage and cohort reuse. Entries give problems × samples or shared histories per policy. Reused cohorts are listed once by scientific role; each model supplies its own Base histories.

Study	Qwen3-1.7B	Qwen3-4B
Source transfer, complete 4k responses	60×4	60×4
Source/shaping shared-prefix comparison	128×4	128×4
Across-problem state pairing	256×4	128×4
Same-question prefix pairing	256×4	128×4
Cross-question pairing, reverse-KL loss	256×4	—
Aligned/Permuted boundary crossover	256×4	128×4
Aligned/Base boundary crossover	256×4	128×4
Public shared-prefix pairing	60×4	60×4
MATH-500, natural full responses	500×2	500×2
Parameter recomposition	128×4	—
Source–budget surface	64×8	—
Separate-cohort continuation	64×2	—

A.3 Prompt construction

The student receives the problem and the following common final-answer instruction:

> Reason independently. Use `\boxed{}` exactly once, only for the final answer; do not box intermediate values.

For all three contextual sources, the teacher uses the same wrapper, changing only the example fields:

> The following worked example is supplied as additional mathematical context while evaluating the student's continuation on the target.
>
> `=== Context Problem Begin ===`
> *[context problem]*
> `=== Context Problem End ===`
> `=== Context Solution Begin ===`
> *[context solution]*
> `=== Context Solution End ===`
>
> `Problem:` *[target problem]*
>
> *[common final-answer instruction]*

The Fixed source repeats the following problem for every training example: "Prove that for every integer n, the number $n^5 - n$ is divisible by 30." Its frozen solution reads:

> It is enough to prove divisibility by 2, 3, and 5. The product of two consecutive integers n(n-1) is even, so nˆ5-n is even. Modulo 3, every integer is congruent to 0, 1, or -1, and xˆ5-x vanishes in all three cases. Modulo 5, Fermat's theorem gives nˆ5 congruent to n. Since 2, 3, and 5 are pairwise coprime, 30 divides nˆ5-n.

The quotation preserves the frozen experimental prompt, including its ASCII exponent syntax and variable-name switch. The target problem and its sampled response prefix are still available to the Fixed teacher; only its *additional worked example* is fixed. The example is absent from deployed student prompts. Prompt-length checks reject over-limit inputs rather than silently truncating the worked solution.

A.4 Training and decoding

All comparisons at a given scale begin from the same frozen Qwen3 initialization (1.7B or 4B). Contextual teachers are evaluated with the student adapter disabled and with teacher targets detached. LoRA updates (Hu et al., 2022) are applied to attention projections and feed-forward projections, while the base weights remain frozen. Table 3 records the paired intervention's training and evaluation configuration. Source and shaping experiments share the 100-step schedule, LoRA rank/scaling, learning rate, completion cap, and decoding distribution; their dataset rows are single

examples. Source/shaping seeds are 41/42/43; pairing seeds are 51/52/53. Every reported trained policy is the fixed step-100 checkpoint.

Table 3: Configuration of the state-pairing experiment. The batch unit is a *pair* of generated examples: four devices with one pair per device and two accumulation steps give 16 examples per full accumulation step; epoch-tail steps may contain fewer examples (Appendix E.7).

Setting	Value
Student / frozen reference	Qwen3-1.7B or Qwen3-4B, thinking enabled
Teacher context	Same frozen mathematical example for every problem
Training examples	98 problems: 49 cross-problem pairs or 98 same-question pairs
Optimizer steps; training seeds	100; 51, 52, 53
LoRA rank / scaling parameter	64 / 128
LoRA target projections	q, k, v, o attention; gate, up, down feed-forward
Learning rate	2×10^{-6}
Devices / pair batch / accumulation	4 / 1 pair per device / 2
Rollout generation / engine synchronization	Generation inside each training-step call; vLLM synchronized each optimizer step
Maximum completion / total length	4,096 / 8,192 tokens during training
Teacher computation	Detached, float32 distribution arithmetic
Distillation objective	Full-vocabulary symmetric JSD; reverse-KL in the objective comparison
Target-strength matching	Recipient-reference full-vocabulary JSD in all pairing arms
Response-position chunk size	32 tokens; no vocabulary truncation
Runtime	BF16, gradient checkpointing, DeepSpeed ZeRO-2
Sampling temperature / top-p / top-k	0.6 / 0.95 / 20
Evaluation history / continuation caps	4,096 / 4,096 tokens
Evaluation histories	Four frozen Base histories per problem
Boundary switch	First single-token `</think>`; exact-prefix re-encoding

The 4B run uses four A800-80GB GPUs; both scales retain the paired-example training configuration in Table 3. The pinned runtime uses PyTorch 2.8.0, Transformers 4.57.1, TRL 0.26.0, Accelerate 1.11.0, DeepSpeed 0.18.2, PEFT 0.17.1, Datasets 3.6.0, and vLLM 0.11.0. Generation uses no presence penalty. Evaluation re-encodes stored token IDs, not decoded and re-tokenized strings. For the state-pairing matrix, history, pre-boundary, and post-boundary random seeds are deterministic functions of problem position and history index, shared across policy regimes. Equal seeds reduce avoidable sampling variation but do not imply identical continuations under different distributions.

A.5 Answer and stopping conventions

Scoring extracts the last balanced `\boxed{...}` expression in the deployable answer region. If a thinking block is open and never closes, a boxed expression inside that block does not count as an exposed answer. With a closed thinking block, scoring uses the region after its closing marker. The fixed mathematical-equivalence checker supports the frozen presentation aliases and normalization rules; multiple boxes and disagreement between first and final boxes are retained as diagnostics rather than used to select a favorable answer.

Natural termination and exhaustion of the token allowance are recorded separately. A history that already stopped is absorbing under all continuation regimes. Among unfinished histories, the policy receives the remaining token allowance; the boundary switch does not grant a second full budget. The outcome record distinguishes a correct valid terminal answer, a wrong valid terminal answer, malformed termination, and budget censoring. Unclosed thinking is additionally recorded and can overlap censoring. Censored trajectories retain their observed budget-limited outcomes and remain in the full-history denominator.

A.6 Operational specification of the controlled comparisons

The implementation is defined by the computational relationships in Table 4. In training, the frozen teacher evaluates the student's sampled prefixes in two conditioning views. In evaluation, the trained policy receives the stored history tokens and a fixed remaining allowance. For a crossover, the

pre-policy emits the closing thinking delimiter; the post-policy then re-encodes the complete token prefix and continues within that same allowance. Same-policy crossovers use the same re-encoding operation. These details preserve the intended comparison across training and evaluation conditions.

Table 4: Computational specifications that preserve the experimental controls.

Operation	Specification
Context construction	Identical wrapper; only the declared reference fields vary.
Teacher evaluation	Frozen initialization, student adapter disabled, shared sampled response prefix.
Direction exchange	Recipient q_0 retained; donor displacement recentered and rescaled under the recipient.
Target calibration	Actual full-vocabulary JSD verified against Equation 9.
Training support	Both candidate targets use the same accepted positions within each sampled batch.
Policy crossover	Actual `</think>` boundary; exact-token re-encoding and a shared remaining budget.
Answer evaluation	Exposed final-answer region, fixed equivalence rule, separate terminal and censoring records.
Statistical aggregation	Histories averaged within problem and seed; paired contrasts use a common Base record.

Experiment records associate each checkpoint with its initialization, training condition, seed, and frozen evaluation histories. Scalar outcome records preserve problem identifiers, correctness, completion status, and token usage; the statistical units and reuse rules are specified in Appendix E.4. The public dataset sources retain their original attribution and terms. Third-party problem text, reference answers, generated responses, and weights are managed separately from the paper's aggregate results.

A.7 Natural MATH-500 evaluation and budget readouts

Section 6.3 evaluates the original cross-question JSD-trained Aligned and StatePermuted checkpoints, each at step 100 and with training seeds 51/52/53, together with one Base policy per scale. All seven policies generate complete responses from the question alone, with thinking enabled and the common final-answer instruction. The matrix uses neither the same-question nor the reverse-KL checkpoints. It introduces no additional parameter updates.

The test set is the fixed 500-question `HuggingFaceH4/MATH-500` subset of MATH (Hendrycks et al., 2021), introduced by Lightman et al. (2024), at revision `6e4ed1a2a79af7d8630a6b768ec859cb5af4d3be`. Screening against the registered training and context examples identifies zero known overlaps; the resulting clean subset equals the full 500-question set. The seven subjects contain 124 Algebra, 38 Counting & Probability, 41 Geometry, 97 Intermediate Algebra, 62 Number Theory, 82 Prealgebra, and 56 Precalculus questions. Levels 1–5 contain 43, 90, 105, 128, and 134 questions. These dataset-defined groups are used for descriptive coverage checks.

Generation uses temperature 0.6, top-p 0.95, top-k 20 and a 32,768-token completion cap. Matching question/sample indices share the same prompt and decoding seed across policies. The prespecified primary is the Aligned–StatePermuted Avg@1 contrast at 32,768 tokens using sample 0 (K1). The completed time-gated extension adds sample 1 to every model–policy–question cell, yielding K2; activation depends on elapsed time and matrix coverage. K2 contains K1. The four-sample extension was not executed. There are $2 \times 7 \times 500 \times 2 = 14{,}000$ distinct generated responses and 56,000 budget readouts.

For a budget $b \in \{4096, 8192, 16384, 32768\}$, success requires natural stopping by b, a closed thinking block, and a correct exposed final boxed answer. Shorter-budget entries are read from the same longest-budget response; generation is neither restarted nor forced to emit an answer at the cutoff. Here the budget begins at the question, whereas shared-prefix experiments allocate a separate continuation allowance after a fixed history. Mean observed tokens are $\min(T, b)$ for generated length T; all responses, including incorrect and censored ones, enter this average.

Table 5: **Source choice changes the learned continuation policy.** Avg@4 (%) on the common 128-problem source/shaping cohort, four model-specific frozen histories per problem. Each trained entry averages three training seeds. Bottom rows give paired differences in percentage points; full intervals appear in Appendix B.1.

Training source or rule	Qwen3-1.7B	Qwen3-4B
Base	44.92	49.61
Target	50.07	52.15
Other	47.98	49.80
Fixed	48.24	50.46
SelfSharp	44.66	49.87
TailSuppress	44.21	47.98
ContextMean	48.76	50.80
Target − Other	+2.08	+2.34
ContextMean − SelfSharp	+4.10	+0.93
ContextMean − TailSuppress	+4.56	+2.82

The primary answer checker is the frozen `math-verify` 0.8.0 implementation with explicit TeX environments and the last complete exposed box. The previous checker is retained for a sensitivity analysis on the same responses (Appendix D.6). Correct, wrong-valid, malformed and censored outcomes partition each budget's sample; unclosed thinking is an overlapping diagnostic. A corrected inclusive boundary-end comparison changes one descriptive boundary flag, while correctness, completion, censoring and token-length readouts retain their original values.

For each scale and budget, samples are averaged within question and training seed before paired resampling of the 500 questions and the three fixed training seeds. Base outcomes are shared by the seed-specific contrasts. This evaluation uses 20,000 crossed bootstrap replicates, NumPy random seed 24601, batch size 256, and linear percentile interpolation. Intervals are pointwise 95%. This prespecified analysis is separate from the shared-prefix resampling implementation in Appendix E.4. The four budgets and cumulative sampling rounds are correlated readouts of the same policies and questions.

B Reference Value and Transfer

This appendix supports Section 4 by separating source-specific learning effects, public-benchmark transfer, and the functional reuse of non-target-trained updates. Each experiment answers a particular aspect of reference value; its cohort, generation budget, and estimator are stated locally. Tables identify the cohort, metric, and interval level for each comparison.

B.1 Source-specific continuation and probability-shaping controls

To unpack the pooled improvement in Section 4.2, the two source/shaping studies evaluate the same 128 questions on different model-specific Base histories. ContextMean averages outcomes from Target, Other, and Fixed; it is an evaluation summary, while ControlCentroid in Appendix B.3 denotes a parameter combination. Table 6 reports individual source effects and the direct contrasts supporting Sections 4.2 and 4.3.

The non-target average, $(O + F)/2$, improves over Base by 3.190 points $[1.172, 5.436]$ on 1.7B and 0.521 $[-0.944, 1.986]$ on 4B. Target–Fixed is 1.823 $[-0.065, 3.971]$ and 1.693 $[-0.391, 4.297]$. Target–Other is positive in all six training-seed comparisons: $1.758/1.367/3.125$ and $3.320/2.148/1.562$. These contrasts describe whole source interventions, which include the example's content, relevance, and length.

B.2 Reference transfer and pairing checks on public benchmarks

This comparison connects the source effects of Section 4.2 and the pairing effects of Section 5.1 to named contest benchmarks. The 60 questions comprise 30 AIME25 and 30 HMMT25 problems. Source transfer samples four full responses per question with a 4,096-token cap. Public pairing in-

Table 6: Source and shaping contrasts, with standardized paired 95% intervals.

	Qwen3-1.7B		Qwen3-4B	
Contrast	Effect	Interval	Effect	Interval
Target − Base	+5.14	$[+2.86, +7.62]$	+2.54	$[+0.72, +4.62]$
Other − Base	+3.06	$[+0.72, +5.66]$	+0.20	$[-1.30, +1.76]$
Fixed − Base	+3.32	$[+1.04, +5.73]$	+0.85	$[-1.04, +2.73]$
SelfSharp − Base	-0.26	$[-2.34, +1.76]$	+0.26	$[-1.37, +1.82]$
TailSuppress − Base	-0.72	$[-2.93, +1.50]$	-1.63	$[-3.26, 0.00]$
ContextMean − Base	+3.84	$[+1.87, +6.01]$	+1.19	$[-0.13, +2.60]$
Target − Other	+2.08	$[+0.13, +4.10]$	+2.34	$[+0.59, +4.43]$
ContextMean − SelfSharp	+4.10	$[+2.21, +6.12]$	+0.93	$[-0.26, +2.15]$
ContextMean − TailSuppress	+4.56	$[+2.80, +6.42]$	+2.82	$[+1.32, +4.45]$

stead continues four frozen Base histories per question for up to another 4,096 tokens. Its policies are the original cross-problem pairing checkpoints. Table 7 keeps these channels separate. The 4B

Table 7: Public evaluation channels and their standardized 95% intervals.

Channel and contrast	Model	Effect [interval]
Full response: Target − Base	4B	$+1.53$ $[-0.83, +4.31]$
Full response: Other − Base	4B	$+0.56$ $[-1.67, +2.50]$
Full response: Fixed − Base	4B	$+0.83$ $[-0.83, +2.64]$
Full response: Target − Other	4B	$+0.97$ $[-1.94, +3.61]$
Shared prefix: Aligned − Base	1.7B	$+0.83$ $[-0.28, +2.08]$
Shared prefix: Aligned − StatePermuted	1.7B	$+0.69$ $[-0.28, +1.81]$
Shared prefix: StatePermuted − Base	1.7B	$+0.14$ $[-0.56, +0.83]$
Shared prefix: Aligned − Base	4B	$+1.81$ $[0.00, +4.17]$
Shared prefix: Aligned − StatePermuted	4B	$+2.08$ $[0.00, +5.28]$
Shared prefix: StatePermuted − Base	4B	-0.28 $[-2.08, +1.25]$

full-response accuracies are 3.750%, 5.278%, 4.306%, and 4.583% for Base, Target, Other, and Fixed. Base has 92.083% unclosed thinking at this cap. The shared-prefix native-pairing differences by training seed are $1.250/0.833/0.000$ for 1.7B and $2.500/0.417/3.333$ for 4B. These public results have the same aggregate ordering as the internal pairing cohorts, with the precision shown in the table.

Table 8: Qwen3-1.7B complete-response source-transfer effects on AIME25/HMMT25, in percentage points. Intervals use 10,000 paired-problem/training-seed bootstrap resamples at 90% coverage.

Contrast	Effect	90% interval
Target − Base	3.056	$[0.833, 5.972]$
Other − Base	2.361	$[0.139, 5.000]$
Fixed − Base	1.806	$[0.139, 3.750]$
Target − Other	0.694	$[-1.111, 3.194]$
Target − Fixed	1.250	$[-0.417, 3.333]$
Other − Fixed	0.556	$[-1.528, 2.778]$
Target − Other, AIME25	0.833	$[-1.667, 4.167]$
Target − Other, HMMT25	0.556	$[-1.389, 3.611]$

Table 8 gives the complementary 1.7B complete-response results and benchmark-specific contrasts. Its three Target–Other seed differences are 0.417, -0.833, and 2.500 points. The pooled non-target gains support the reuse of worked references on these contest questions; the source-specific shared-prefix test in Appendix B.1 examines their value from controlled histories. The 90% intervals here retain this evaluation's prespecified analysis; Appendix E.4 distinguishes that specification from the standardized 95% summaries.

B.3 FUNCTIONAL VALUE OF COMBINED NON-TARGET SOURCE UPDATES

Section 4.2 establishes that training with non-target references can yield useful policies. We complement that behavioral result by asking whether the resulting parameter changes remain useful when the Other and Fixed checkpoints are combined. This test measures functional reuse of source-trained updates, alongside the direct source-specific continuation tests.

For each source-training seed, write the source checkpoints as θ_T, θ_O, and θ_F, with initialization θ_0. The arithmetic construction is

$$\theta_{\text{cent}} = \tfrac{1}{2}(\theta_O + \theta_F), \qquad \theta_{\text{dev}} = \theta_0 + \theta_T - \theta_{\text{cent}}, \qquad \theta_{\text{recon}} = \theta_{\text{cent}} + \theta_{\text{dev}} - \theta_0. \tag{11}$$

In exact arithmetic, the reconstruction equals θ_T. The components define executable parameter perturbations. Because the deviation depends algebraically on the centroid, this construction evaluates functional parameter combinations; its interpretation concerns the utility of those combinations.

Table 9: Qwen3-1.7B functional recomposition on the independent 128-problem cohort. All effects are percentage points with 90% intervals. "Original" and "reconstructed" distinguish storage paths as well as the arithmetic construction.

Contrast	Effect	90% interval
Original Target − Base	7.03	$[4.49, 9.77]$
Reconstructed Target − Base	7.62	$[4.95, 10.48]$
Reconstructed − Original Target	0.59	$[-1.69, 2.80]$
Control centroid − Base	5.60	$[2.86, 8.46]$
Target deviation − Base	2.08	$[0.33, 3.97]$
Reconstructed Target − Control centroid	2.02	$[-0.46, 4.69]$
Arithmetic nonadditivity	−0.07	$[-3.32, 3.19]$

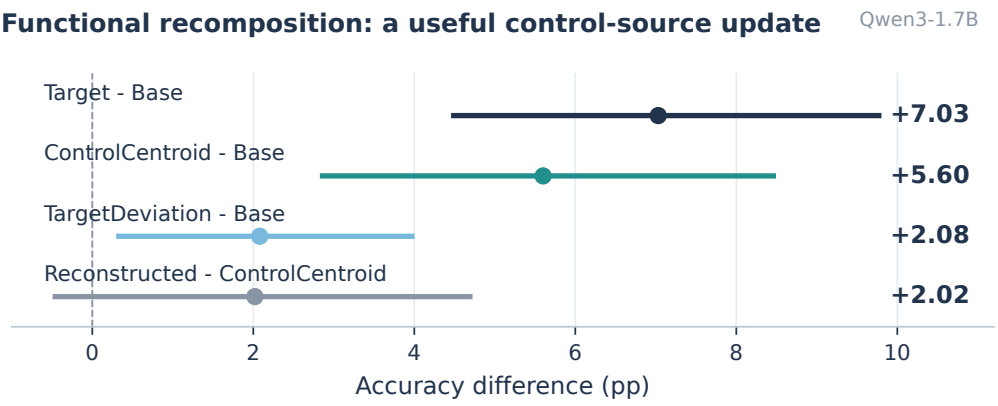

Figure 6: **Non-target source updates retain functional value under recomposition.** Results on the separate 128-problem recomposition cohort; intervals are 90%. The source checkpoints have been materialized in BF16, so this test complements rather than replaces the adapter-native precision check.

B.4 NUMERICAL PRECISION AND INFERENCE BUDGET IN PARAMETER RECOMPOSITION

The materialized recomposition uses full checkpoints in BF16. Its rounding discrepancy is large relative to the small intended update: approximately 77% of update L_2 for the centroid and 40–41% for the deviation and reconstruction. The reconstruction comparison has an interval spanning zero with no prespecified equivalence margin. The adapter-native construction below complements this materialized result by expressing the intended update before full-checkpoint rounding. These two representations clarify how the measured functional effects depend on arithmetic precision.

For each adapted weight matrix, let $A_X \in \mathbb{R}^{64\times d_{\text{in}}}$ and $B_X \in \mathbb{R}^{d_{\text{out}}\times 64}$ be its LoRA factors for source $X \in \{T, O, F\}$. Rank 64 and scaling 128 give the effective update $\Delta_X = 2B_X A_X$. The

native centroid uses

$$A_{\text{cent}} = \begin{bmatrix} A_O \\ A_F \end{bmatrix}, \quad B_{\text{cent}} = [B_O \quad B_F], \quad \Delta_{\text{cent}} = B_{\text{cent}} A_{\text{cent}} = \tfrac{1}{2}(\Delta_O + \Delta_F), \tag{12}$$

with rank/scaling 128/128. The native deviation uses

$$A_{\text{dev}} = \begin{bmatrix} A_T \\ A_O \\ A_F \\ 0 \end{bmatrix}, \quad B_{\text{dev}} = [2B_T \quad -B_O \quad -B_F \quad 0], \quad \Delta_{\text{dev}} = \Delta_T - \Delta_{\text{cent}}, \tag{13}$$

with rank/scaling 256/256. The last 64 coordinates are zero padding for runtime rank support, not an additional learned component.

To examine source-update reuse over inference budgets, the source–budget study evaluates 25 policy configurations on a separate cohort of 64 problems with eight responses each. Each of the 12,800 responses is generated once up to 8,192 tokens and scored at 2,048, 4,096, and 8,192 tokens. The three budgets are nested readouts of these responses. The main adapter-native contrast, Target minus ControlCentroid in Pass@8 at 8,192 tokens, is -1.04 points with 90% interval $[-5.21, 3.13]$. The comparison describes this cohort, precision path, budget, and sampling metric; the source-specific continuation tests in Section 4.2 use their own common-history protocol.

The recomposition and budget experiments extend the source-content question in two directions: whether non-target-trained updates remain useful after parameter combination, and how their observed performance depends on the numerical representation and available inference budget. The shared-prefix source comparison then evaluates complete trained policies from matched visible histories, giving the primary readout for reference value in the main text.

B.5 Continuation gains on a separate problem cohort

An additional 64-problem evaluation continues two frozen Base histories per problem, reusing the source–budget question set with a distinct continuation intervention. ContextMean improves Avg@2 by 4.43 points, with a 90% interval $[1.13, 7.81]$. Valid completion increases by 5.73 points, with an equal reduction in censoring. The separate 128-problem 1.7B source/shaping cohort yields a 3.84-point Avg@4 improvement. Together, these comparisons support a learned continuation benefit across two question cohorts and two per-question sampling counts. They complement Section 4.2's source-specific analysis; the state-pairing intervention is evaluated on its own cohorts in Appendix C.

C State-Specific Guidance and Objective Robustness

The experiments in Section 5 progressively examine the relation between a teacher direction and the state that produced it. Cross-question exchange tests the complete question–prefix state; same-question exchange holds the question fixed; changing the training loss tests the same target-assignment intervention under another optimization objective. This appendix provides seed-resolved evidence, supplementary metrics, and the precise relationships among these comparisons.

C.1 Cross-question pairing across model scales

Section 5.1 exchanges context-induced directions across within-family question pairs while retaining each recipient prediction. Table 10 reports the three paired training-seed estimates at each scale alongside their cohort-specific pooled contrasts.

The point estimates and seed differences support the same native-pairing ordering at both scales. The table retains the prespecified 95% intervals, while Table 1 uses the standardized implementation defined in Appendix E.4; they are analyses of the same endpoint observations.

The seed-specific Aligned–Base differences are $+4.10$, $+3.13$, and $+1.95$ points; StatePermuted–Base gives $+1.17$, -0.39, and -0.20 points. After histories and seeds are averaged within problem, Aligned–StatePermuted has 29 positive, 8 negative, and 91 tied problem-level differences. The corresponding Aligned–Base counts are 28, 10, and 90. The cohort-level intervals use problems and the finite three-seed set as their sampling units.

Table 10: Aligned minus StatePermuted shared-prefix accuracy under the prespecified analysis (percentage points). Individual seed estimates are shown separately from the cohort-level 95% intervals. All six paired seed contrasts are positive.

Training seed	Qwen3-1.7B	Qwen3-4B
51	+4.88	+2.93
52	+4.79	+3.52
53	+2.93	+2.15
Pooled	+4.20	+2.86
95% interval	$[+2.41, +6.02]$	$[+0.46, +5.40]$

Table 11: Cross-question 4B pairwise effects (percentage points). Avg@4 and supplementary Pass@4 use the original prespecified 95% crossed-bootstrap intervals. Each contrast compares policies on the same question/history cohort.

	Avg@4		Pass@4	
Contrast	Effect	95% interval	Effect	95% interval
Aligned − Base	+3.06	$[+0.59, +5.53]$	+3.91	$[+0.26, +7.81]$
Aligned − StatePermuted	+2.86	$[+0.46, +5.40]$	+1.82	$[-1.56, +5.47]$
StatePermuted − Base	+0.20	$[-2.60, +2.93]$	+2.08	$[-2.08, +6.25]$

C.2 Same-question prefix specificity

To hold question identity fixed in Section 5.2, the experiment forms 98 same-question pairs, each preserving question and fixed-context identity, derived from the original 49 cross-problem pairs. Pair construction precedes task-model outputs. Each pair independently samples two trajectories. Positions enter the intervention after the visible prefixes diverge and while both trajectories occupy the same pre-token generation phase. Both arms use the same within-batch eligibility rule and recipient-reference matching. They generate their own on-policy trajectories during training.

At 1.7B, the six independently trained policies yield 6,144 evaluation endpoints on 256 questions with four shared Base histories each. Adding the 1,024 shared Base endpoints gives 7,168 records. Native pairing exceeds exchange by $3.516/2.734/3.125$ points across seeds; native minus Base is $2.930/2.734/2.246$, and exchanged minus Base is $-0.586/0.000/-0.879$. Averaging histories and seeds within problem gives 64 positive, 24 negative, and 168 tied native–exchanged differences. Native and exchanged policies produce 1,299 and 1,203 correct endpoints, respectively, out of 3,072 each.

The exploratory Pass@4 native–exchanged difference is 3.516 points $[0.911, 6.250]$; native–Base is 2.995 $[-0.130, 6.380]$. Avg@4 is the primary outcome; Pass@4 describes successful attempts within a four-history group. Each pairing scope has its own matched training organization and on-policy trajectories.

Shared question identity across scales. Both scales use the same 98 training questions and same-question construction. For 4B, six independently trained policies are evaluated on the disjoint 128-question cohort. Its six cells contribute 3,072 new endpoints, combined with 512 shared Base endpoints. IntraAligned produces 772 correct outcomes out of 1,536, versus 727 for IntraPermuted and 243 of 512 for Base. The paired native–exchanged effect is $+2.930$ $[0.651, 5.143]$ points; native–Base is $+2.799$ $[0.195, 5.599]$ and exchanged–Base is -0.130 $[-2.995, 2.865]$. Together with the 1.7B comparison, the positive 4B effect identifies the partial solution as a functional component of useful guidance, with question and reference fixed within each training pair.

C.3 Sensitivity to the optimization objective

Section 5.3 tests whether the pairing advantage persists when the student learns through a different loss. We repeat the cross-question experiment with $D_{\mathrm{KL}}(p_\theta \| q)$ as the student loss, using the original 49 training pairs, fixed proof context, phase eligibility, and recipient-reference JSD calibration. The

Table 12: Same-question paired differences by training seed (percentage points). Each model uses its own fixed evaluation cohort; Base is shared across the three seeds.

Model	Contrast	Seed 51	Seed 52	Seed 53
1.7B	Native − Exchanged	+3.516	+2.734	+3.125
1.7B	Native − Base	+2.930	+2.734	+2.246
1.7B	Exchanged − Base	-0.586	0.000	-0.879
4B	Native − Exchanged	+4.102	+2.734	+1.953
4B	Native − Base	+2.344	+4.102	+1.953
4B	Exchanged − Base	-1.758	+1.367	0.000

target is detached, and both loss and target arithmetic use the full vocabulary in float32 at temperature one. For one eligible state, the reverse-KL loss and its logit gradient are

$$\ell = \sum_a p_a(\log p_a - \log q_a), \qquad \frac{\partial \ell}{\partial z_a} = p_a(\log p_a - \log q_a - \ell), \tag{14}$$

where $p = \mathrm{softmax}(z)$, $p_a = p_\theta(a \mid s)$, and $q_a = q(a \mid s)$ for the detached target $q \in \{q_A, q_P\}$ at the same state s. The common response mask is applied before normalization over participating positions. Thus the experiment varies the training objective while keeping the target-matching geometry in Equation 9 fixed.

Six new step-100 policies are evaluated on the same 256 questions and four Base histories per question as the JSD comparison. The new cells contribute 6,144 endpoints; the self-contained matrix adds 1,024 shared Base endpoints. Native and exchanged policies yield 1,315 and 1,191 correct outcomes out of 3,072 each. Native–exchanged is $+4.036$ $[2.083, 6.055]$ points, native–Base is $+3.158$ $[1.432, 4.980]$, and exchanged–Base is -0.879 $[-3.060, 1.172]$.

Table 13: Reverse-KL cross-question comparisons by training seed (percentage points).

Contrast	Seed 51	Seed 52	Seed 53
Native − Exchanged	+4.980	+2.539	+4.590
Native − Base	+2.539	+2.930	+4.004
Exchanged − Base	-2.441	+0.391	-0.586

For comparison, the standardized JSD native–exchanged effect on these histories is $+4.199$ $[2.344, 6.055]$ points. An exploratory paired difference of the two contrasts (reverse KL minus JSD) is -0.163 $[-2.669, 2.441]$. The positive pairing effects under both losses extend the functional state-dependence result across optimization objectives. The difference-of-contrasts reports their observed difference; an equivalence margin was not specified for this exploratory comparison.

D Reasoning Stages and Completion Dynamics

Section 6 connects matched guidance to successful reasoning completion. We expand that connection through three complementary readouts: deployment-stage policy crossovers, full-cohort terminal outcomes, and the evolution of those outcomes within a fixed token budget. These comparisons retain their common-history denominators and explicitly distinguish the policy contrasts they measure.

D.1 Deployment-stage comparisons and conditional answer-policy effects

To support Section 6.1, we report both the positive training-gain comparison and the native-versus-exchanged comparison using a common two-stage notation. Let the first letter indicate the policy before the actual thinking boundary and the second the policy after it. In the A/B experiment, B denotes Base; in the A/P experiment, P denotes StatePermuted. The closing delimiter is emitted by the first policy. Prefix token hashes, lengths, and boundary states agree within each fixed-pre-policy

Table 14: Positive training-gain crossovers: A =Aligned, B =Base. Standardized 95% intervals.

	Qwen3-1.7B		Qwen3-4B	
Contrast	Effect	Interval	Effect	Interval
$AA - BB$	+3.06	$[+1.30, +4.95]$	+3.06	$[+0.59, +5.60]$
$AB - BB$	+2.80	$[+1.11, +4.59]$	+2.86	$[+0.46, +5.34]$
$BA - BB$	+0.55	$[0.00, +1.20]$	+0.46	$[-0.13, +1.17]$
$AA - AB$	+0.26	$[-0.20, +0.78]$	+0.20	$[-0.39, +0.91]$
$AA - BA$	+2.51	$[+0.78, +4.36]$	+2.60	$[0.00, +5.21]$

Table 15: **Matched guidance increases correct completion.** Native minus exchanged-direction rates (percentage points) on each full shared-prefix cohort. Valid equals correct plus wrong-valid; censoring is token-cap exhaustion. Target matching uses JSD; the loss column identifies the optimization objective.

Exchange	Model	Loss	Correct	Valid	Wrong-valid	Censored
Across questions	1.7B	JSD	+4.20	+5.21	+1.01	-5.21
Across questions	4B	JSD	+2.86	+2.80	-0.07	-2.80
Same question	1.7B	JSD	+3.13	+4.36	+1.24	-4.36
Same question	4B	JSD	+2.93	+3.13	+0.20	-3.13
Across questions	1.7B	Reverse KL	+4.04	+5.01	+0.98	-5.01

contrast. Naturally absorbed histories are retained in the whole-cohort result. The thinking-only retention statistic is $(V_{AB} - V_{BB})/(V_{AA} - V_{BB})$, equal to 91.489% and 93.617%. Averaging the two conditional decomposition paths gives pre-/post-boundary contributions of $2.653/0.407$ and $2.734/0.326$ points. The corresponding pre shares are 86.7% and 89.4%. Retained gain under a deployment rule and the symmetric decomposition are different summaries of the same 2×2 matrix.

For the A/P comparison at 4B, the standardized results are $AP - PP = 2.539$ $[0.000, 5.078]$, $AA - PA = 2.799$ $[0.326, 5.404]$, $AA - AP = 0.326$ $[-0.391, 1.172]$, and $PA - PP = 0.065$ $[-0.586, 0.781]$. At 1.7B the two pre effects are 3.74–3.78 points and the two post effects 0.42–0.46; these original conditional results remain in the registered profile. Here the ranges denote two conditional point estimates, not confidence intervals.

Actual answer-policy calls. A conditional secondary analysis uses keys $(\text{training seed}, \text{problem}, \text{history})$ and retains histories in which the compared policies actually call the post-boundary policy. Absorbed histories are excluded. Problems must have eligible histories in all three seeds. For $AA - AB$, the effects are 0.913 $[-0.447, 2.685]$ on 149 common problems and 1.556 $[-0.333, 4.778]$ on 75. For $BA - BB$, they are 1.617 $[0.000, 3.639]$ on 158 problems and 2.092 $[-0.108, 5.519]$ on 77. These are selected, fixed-pre-policy populations; their coverage and estimates complement the full-cohort deployment effects.

Conditioning on actual post-policy invocation changes the evaluation population. The full-cohort contrasts quantify deployment effects; the selected comparisons quantify answer-policy differences within their stated calling populations. Seed identity is retained throughout selection and paired inference (Appendix E.4).

D.2 Correctness, completion, and censoring on the full cohorts

To interpret the accuracy changes in Section 6.2, we jointly report correctness, valid completion, and unfinished trajectories. Correct and wrong-valid outcomes partition valid completion. Open thinking can overlap token-cap censoring. All rates below keep the full common-history denominator. The conditional correct/valid ratio describes each policy's own completed set; primary accuracy is evaluated on all histories. For cross-problem 1.7B, native–exchanged valid completion is 5.208 $[3.158, 7.259]$, wrong-valid completion 1.009 $[0.195, 1.855]$, and open thinking -4.818 $[-6.803, -2.865]$. For same-question pairing, the corresponding effects are

Table 16: Absolute outcomes on the 256-problem pairing cohort. Cross-question JSD, same-question JSD, and cross-question reverse-KL arms are distinct training experiments. Length is mean additional tokens.

Policy	Correct	Valid	Wrong-valid	Censored	Open	Length
Base	39.65	44.34	4.69	55.66	49.71	2862.0
Cross: Native	42.71	48.21	5.50	51.79	46.13	2777.7
Cross: Exchanged	38.51	43.00	4.49	57.00	50.94	2869.0
Same: Native	42.29	47.88	5.60	52.12	46.52	2772.0
Same: Exchanged	39.16	43.52	4.36	56.48	51.27	2868.8
Reverse KL: Native	42.81	48.27	5.47	51.73	46.45	2776.3
Reverse KL: Exchanged	38.77	43.26	4.49	56.74	51.11	2871.4

4.362 $[2.799, 6.055]$, 1.237 $[0.488, 2.051]$, and -4.753 $[-6.641, -2.832]$. Table 19 reports cross-question 4B absolute outcomes; Table 23 gives the corresponding behavior intervals.

Table 17: Source/shaping completion outcomes on the common 128-problem cohort. Percentages.

Model	Policy	Correct	Valid	Censored	Open
1.7B	Base	44.92	48.44	51.56	45.70
	ContextMean	48.76	53.65	46.35	40.39
	SelfSharp	44.66	48.83	51.17	44.99
	TailSuppress	44.21	47.72	52.28	46.68
4B	Base	49.61	52.54	47.46	42.58
	ContextMean	50.80	54.14	45.86	40.32
	SelfSharp	49.87	52.80	47.20	41.73
	TailSuppress	47.98	50.91	49.09	43.75

Table 18: Same-question 4B outcomes on the 128-problem cohort, using the same histories as the cross-question experiment. Rates are percentages and lengths are mean continuation tokens.

Policy	Correct	Valid	Wrong-valid	Censored	Open	Length
Base	47.46	47.66	0.20	52.34	45.31	2789.7
IntraAligned	50.26	50.46	0.20	49.54	43.36	2703.8
IntraPermuted	47.33	47.33	0.00	52.67	46.61	2790.5

For the same-question 4B comparison, native–exchanged valid completion is $+3.125$ $[0.911, 5.339]$ points, wrong-valid completion is $+0.195$ $[0.000, 0.586]$, and open thinking is -3.255 $[-5.729, -0.977]$. Mean continuation length falls by 86.740 tokens. For the reverse-KL comparison, valid completion rises by 5.013 $[2.930, 7.129]$ points, wrong-valid completion by 0.977 $[0.293, 1.693]$, and open thinking falls by 4.655 $[2.572, 6.771]$ points. Its mean continuation length falls by 95.150 tokens. Length changes are descriptive point estimates. Correct, wrong-valid, and censored frequencies retain the common all-history denominator.

Cross-question 4B cohort and initial histories. The 128-question cohort supplies 512 frozen histories: 69 are already terminal, 405 retain open thinking, and 38 have closed thinking without a terminal answer. Base and all trained policies use these same histories. Table 19 supplies the cross-question results on this denominator; the same-question comparison above reuses it.

D.3 Budget-resolved continuation and boundary transitions

The source/shaping windows read outcomes at 256, 512, 1,024, 2,048, and 4,096 new tokens from the same long continuations. The 4B ContextMean–Base differences are $0.217/0.239/0.803/2.843/1.194$ points. Their original 90% intervals are $[0.022, 0.456]$, $[-0.282, 0.716]$, $[-0.065, 1.736]$, $[1.280, 4.470]$, and $[0.087, 2.344]$, respectively. The 1.7B windows are reported in Table 20. These measurements describe the trajectory's time course under a

Table 19: Qwen3-4B absolute continuation outcomes. Aligned and StatePermuted entries average their three separately evaluated training seeds. Probability measures are percentages; length is the full-cohort mean additional token count.

Outcome	Base	Aligned	StatePermuted
Correct terminal / Avg@4	47.46	50.52	47.66
Pass@4	59.38	63.28	61.46
Valid completion	47.66	50.65	47.85
Right-censoring	52.34	49.35	52.15
Unclosed thinking	45.31	42.25	46.29
Boundary crossed during continuation	33.79	36.85	32.81
Mean continuation tokens	2789.7	2699.9	2779.6

fixed maximum budget. Figure 4(c)–(d) shows their point-estimate curves. The source curve for 1.7B uses the recorded two-decimal summaries; 4B source and pairing curves use the recorded floating-point means. Each table retains its stated interval coverage; the main-text curves show the corresponding point estimates.

Table 20: Correct-terminal gains over additional continuation budget on the shaping cohort (percentage points; rounded profile summaries).

Model	Contrast	256	512	1,024	2,048	4,096
1.7B	Context − Base	0.13	0.13	0.96	2.89	3.84
1.7B	Context − SelfSharp	0.00	0.26	1.09	2.17	4.10

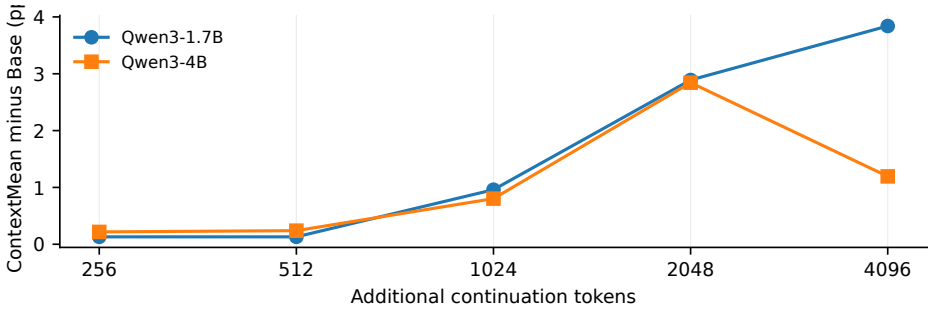

Figure 7: **Contextual gains vary over the continuation budget.** Points are observed windows of the same generated responses; connecting segments are a visual guide. Each model supplies its own histories for the same source/shaping problem IDs.

Boundary occupancy includes histories already past the reasoning boundary. Their contribution is identical across policy regimes, so its paired contrast equals the change in new crossings during continuation. Duration summaries retain all histories, assigning zero pre-boundary continuation tokens to those already terminated or past the boundary. For open histories, pre-boundary time is capped by the continuation budget. These definitions describe the cohort's compute allocation and boundary progress with the same denominator as accuracy.

These measurements complement the policy crossovers in two ways. The crossovers locate where changing the deployed policy affects its outcome; the windowed readouts show when differences in boundary crossing and correct termination emerge. Joint endpoint accounting keeps correct, wrong-valid, and censored outcomes visible, connecting the local supervision intervention to the full-trajectory consequences emphasized in Section 6.

Table 21: Additional native–exchanged correct-terminal gaps over continuation time (percentage points). Each entry is an offline window of the same generated responses; no new decoding or resampling is performed for the window.

Comparison	256	512	1,024	2,048	4,096
Same question, 4B	-0.195	-0.195	+1.497	+2.539	+2.930
Cross-question reverse KL, 1.7B	+0.033	+0.326	+0.911	+2.051	+4.036

Table 22: Budget-resolved 4B Aligned–StatePermuted occupancy differences (percentage points). All five readouts reuse the same generated continuations; intervals are pointwise 95% intervals.

	Boundary occupancy		Correct terminal occupancy	
Additional tokens	Effect	95% interval	Effect	95% interval
256	+1.24	$[+0.07, +2.47]$	-0.13	$[-0.65, +0.33]$
512	+1.11	$[-0.20, +2.73]$	+0.07	$[-0.39, +0.59]$
1,024	+1.95	$[+0.13, +3.91]$	+1.43	$[+0.07, +2.86]$
2,048	+2.41	$[+0.33, +4.69]$	+2.15	$[+0.07, +4.23]$
4,096	+4.04	$[+1.69, +6.58]$	+2.86	$[+0.46, +5.40]$

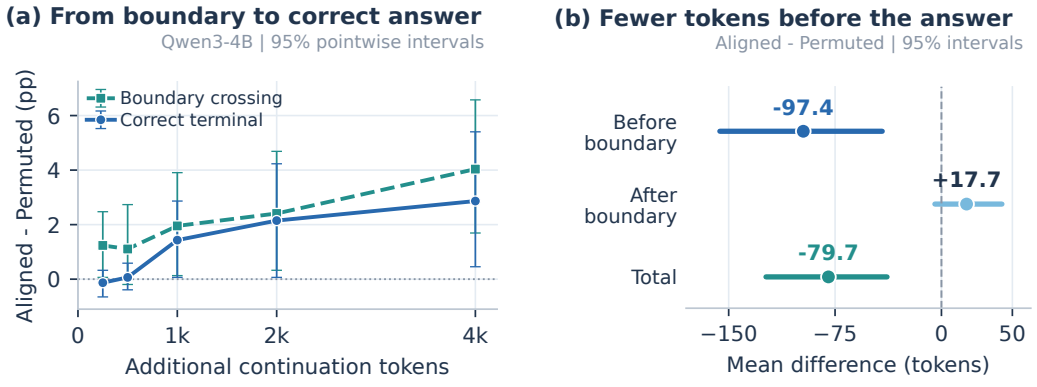

Figure 8: **Boundary progress and token allocation in cross-question 4B pairing.** Native minus exchanged-direction effects on the 128-problem cohort. Left: boundary and correct-terminal occupancy; right: pre-boundary, post-boundary, and total continuation length. Intervals use the original prespecified 95% analysis.

Table 23: **The pairing advantage is expressed in boundary crossing, completion, and token allocation.** Every row compares Aligned with StatePermuted on the 128-problem Qwen3-4B cohort and four shared histories per problem. Probability differences are percentage points; duration differences are tokens. All intervals are 95%.

Outcome	Difference	95% interval
Correct terminal answer	$+2.86$	$[+0.46, +5.40]$
Crosses reasoning–answer boundary	$+4.04$	$[+1.69, +6.58]$
Valid completion	$+2.80$	$[+0.39, +5.34]$
Right-censoring	-2.80	$[-5.34, -0.39]$
Wrong-valid completion	-0.07	$[-0.46, +0.39]$
Pre-boundary continuation tokens	-97.4	$[-156.2, -41.7]$
Post-boundary continuation tokens	$+17.7$	$[-4.6, +42.6]$
Total continuation tokens	-79.7	$[-123.7, -38.4]$

D.4 Full solutions from the question: accuracy across budgets

This analysis extends Section 6.3 from the fixed-history channel to natural complete generation. Tables 24 and 25 report all prespecified budgets, including the one-sample 32k primary endpoint. Aligned–Permuted differences at that primary endpoint are -0.93 points for 1.7B and -0.07 for 4B, with intervals $[-3.00, 1.07]$ and $[-1.20, 1.07]$. The complete two-sample curves localize the positive accuracy differences to 4k–8k. Tables 26 and 27 provide the corresponding Base-relative and training-seed comparisons.

At 4k, both single-sample rounds favor Aligned separately: sample 0 gives 4.27/5.40 points and sample 1 gives 4.33/4.20 across the two scales. All three training-seed contrasts in each round are positive. These repeat decodings characterize sampling consistency for the same trained policies. In the cumulative two-sample matrix, all seven subject groups and all five difficulty levels have positive 4k pairing point estimates; these subgroup summaries are descriptive. The overall, Base-relative and complete budget results remain the inferential comparisons.

Table 24: Prespecified one-sample evaluation (K1). MATH-500 full-response accuracy (%) and paired Aligned minus Permuted differences (pp). Every row uses all 500 questions; 95% intervals are pointwise.

Model	Budget	Base	Aligned	Permuted	$A - P$	95% interval
Qwen3-1.7B	4k	64.00	67.80	63.53	+4.27	$[+1.87, +6.67]$
	8k	81.20	82.47	80.80	+1.67	$[-0.80, +4.13]$
	16k	89.20	88.60	88.67	-0.07	$[-2.13, +2.07]$
	32k	91.00	89.27	90.20	-0.93	$[-3.00, +1.07]$
Qwen3-4B	4k	62.80	68.73	63.33	+5.40	$[+3.20, +7.80]$
	8k	85.60	87.40	85.07	+2.33	$[+0.40, +4.40]$
	16k	93.40	94.20	93.60	+0.60	$[-1.00, +2.20]$
	32k	95.40	95.80	95.87	-0.07	$[-1.20, +1.07]$

D.5 Generation cost and completion behavior

Table 28 reports the terminal partitions, token use, and Pass@2 for every budget. At 4k, Aligned–Permuted valid-completion differences are 4.27 and 4.93 points, alongside correctness differences of 4.30 and 4.80. By 32k, valid completion is 99.4–99.83% across policies and censoring is 0.17–0.60%. Thus, larger allowances give the comparison policies additional opportunities to finish solutions, explaining the observed convergence in success-by-budget curves at the behavioral level.

The Aligned policies use 4,286 and 4,189 generated tokens on average under the 32k cap, versus 4,900 and 4,688 for Permuted. These means include all answers; the respective 12.53% and 10.65% reductions describe generation length. They do not imply the same wall-clock or FLOP reduction. Table 29 gives the paired token differences and an exploratory budget difference-of-contrasts. The latter directly compares each question/seed's 4k and 32k pairing effects, yielding 5.23 and 4.83

Table 25: Complete two-sample evaluation (K2), including K1. MATH-500 full-response accuracy (%) and paired Aligned minus Permuted differences (pp). Every row uses all 500 questions; 95% intervals are pointwise.

Model	Budget	Base	Aligned	Permuted	$A - P$	95% interval
Qwen3-1.7B	4k	63.40	68.23	63.93	+4.30	$[+2.70, +5.97]$
	8k	80.40	82.47	80.77	+1.70	$[+0.20, +3.20]$
	16k	88.60	88.60	88.73	-0.13	$[-1.40, +1.17]$
	32k	90.10	89.57	90.50	-0.93	$[-2.20, +0.30]$
Qwen3-4B	4k	62.80	68.30	63.50	+4.80	$[+3.27, +6.50]$
	8k	85.30	87.70	85.50	+2.20	$[+0.70, +3.73]$
	16k	93.80	94.27	93.93	+0.33	$[-0.90, +1.50]$
	32k	96.00	96.10	96.13	-0.03	$[-0.83, +0.70]$

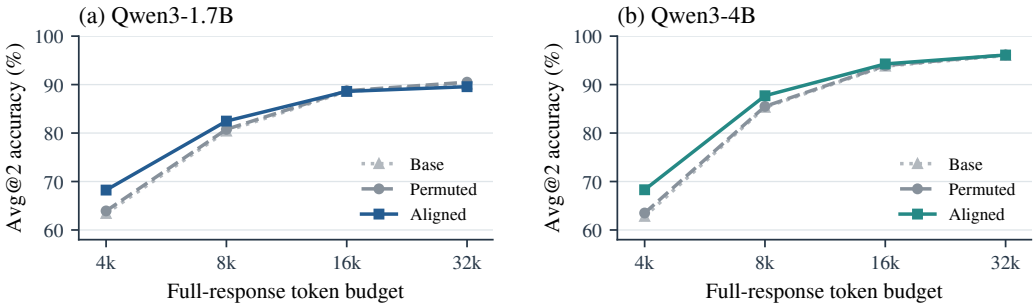

Figure 9: **The full-response accuracy curves approach one another at larger budgets.** MATH-500 Avg@2 uses all 500 questions and the completed two-sample matrix. Each marker is a measured success-by-budget readout of the same 32k-capped responses. Connecting segments guide the eye; complete pointwise contrast intervals are in Tables 25 and 26.

Table 26: Base-relative MATH-500 contrasts (pp). Each interval uses paired problems and the three fixed training seeds. K2 is the cumulative two-sample matrix.

Matrix	Model	Budget	$A - B$	95% interval	$P - B$	95% interval
K1	1.7B	4k	+3.80	$[+1.07, +6.60]$	-0.47	$[-3.13, +2.20]$
		8k	+1.27	$[-1.07, +3.60]$	-0.40	$[-2.80, +2.00]$
		16k	-0.60	$[-2.67, +1.47]$	-0.53	$[-2.73, +1.67]$
		32k	-1.73	$[-3.73, +0.20]$	-0.80	$[-2.87, +1.20]$
K1	4B	4k	+5.93	$[+3.33, +8.73]$	+0.53	$[-2.00, +3.07]$
		8k	+1.80	$[-0.27, +4.00]$	-0.53	$[-2.60, +1.60]$
		16k	+0.80	$[-0.73, +2.40]$	+0.20	$[-1.47, +1.87]$
		32k	+0.40	$[-1.13, +2.00]$	+0.47	$[-0.87, +1.87]$
K2	1.7B	4k	+4.83	$[+2.80, +6.93]$	+0.53	$[-1.33, +2.40]$
		8k	+2.07	$[+0.37, +3.80]$	+0.37	$[-1.40, +2.17]$
		16k	+0.00	$[-1.53, +1.53]$	+0.13	$[-1.43, +1.70]$
		32k	-0.53	$[-1.97, +0.87]$	+0.40	$[-1.00, +1.80]$
K2	4B	4k	+5.50	$[+3.60, +7.53]$	+0.70	$[-1.10, +2.47]$
		8k	+2.40	$[+0.97, +3.90]$	+0.20	$[-1.30, +1.70]$
		16k	+0.47	$[-0.63, +1.60]$	+0.13	$[-0.97, +1.33]$
		32k	+0.10	$[-0.90, +1.10]$	+0.13	$[-0.77, +1.10]$

Table 27: Aligned minus Permuted MATH-500 differences by training seed (pp). Each K1 value uses sample 0; each K2 value averages samples 0 and 1 on the same 500 questions and checkpoints.

Matrix	Model	Budget	Seed 51	Seed 52	Seed 53
K1	Qwen3-1.7B	4k	+3.40	+5.20	+4.20
		8k	+0.20	+3.00	+1.80
		16k	-1.00	+1.00	-0.20
		32k	-2.00	-0.60	-0.20
K1	Qwen3-4B	4k	+5.60	+6.40	+4.20
		8k	+1.20	+2.80	+3.00
		16k	-0.40	+1.00	+1.20
		32k	+0.00	-0.40	+0.20
K2	Qwen3-1.7B	4k	+4.30	+4.00	+4.60
		8k	+1.70	+2.00	+1.40
		16k	-0.10	+0.00	-0.30
		32k	-1.00	-1.20	-0.60
K2	Qwen3-4B	4k	+4.80	+5.50	+4.10
		8k	+1.20	+3.00	+2.40
		16k	-0.60	+0.90	+0.70
		32k	-0.40	+0.00	+0.30

points. Its intervals assess attenuation directly, rather than comparing significance labels at separate budgets.

For 1.7B, the two-sample 32k pairing differences are negative in all three seeds ($-1.00/-1.20/-0.60$ points), while 4B is nearly tied on average. These estimates preserve the potential trade-off between finishing sooner and eventual answer quality. The intervals characterize the precision of the accuracy differences; no non-inferiority or equivalence margin was prespecified. We therefore interpret the combined evidence as improved correct completion at shorter budgets with lower token use, rather than certified equality of long-budget accuracy.

D.6 Answer-checking and final-answer timing sensitivity

Applying both recorded checkers to the same generated responses yields 4.50–5.03% answer-level disagreement at 32k across model–policy groups. The frozen explicit-TeX checker defines the primary results; legacy-checker contrasts provide a sensitivity analysis of the same predictions. Under the legacy checker, 4k Aligned–Permuted gains are 4.50 points $[2.90, 6.13]$ for 1.7B and 4.57 $[3.10, 6.13]$ for 4B. At 32k both intervals include zero. The short-budget pattern is therefore also present under the previous scoring rule. The 1.7B legacy 8k lower endpoint is exactly zero. Scalar scores support this numerical comparison; determining which individual disagreement is mathematically correct requires the corresponding answer text.

A separate final-answer timing check counts responses whose final answer is correct at 32k, whose final exposed box is complete by 4k, and whose natural stop occurs after 4k. Exactly three of the 14,000 responses meet all three conditions. This count is small relative to the 4k correctness gaps and helps distinguish the observed completion behavior from delays after an already completed final box. It concerns the final exposed answer, rather than every provisional value mentioned within reasoning.

The token intervals, direct budget interaction, legacy-checker contrasts and final-box timing checks supplement the prespecified accuracy analysis. Together with the full two-round and four-budget tables, they connect the controlled pairing intervention to solution completion and measured generation cost while retaining its long-budget behavior.

Table 28: Full-response terminal outcomes and token use on MATH-500 (K2). Correct, valid, wrong-valid (W-valid), censored and open-thinking (Open) rates are percentages. Valid comprises correct and wrong-valid completions; all malformed-terminal rates are zero. Mean tokens count actual generation observed up to each budget.

Model	Cap	Policy	Correct	Valid	W-valid	Cens.	Open	Tokens	Pass@2
1.7B	4k	Base	63.40	64.20	0.80	35.80	29.70	2837.4	67.80
		Aligned	68.23	69.20	0.97	30.80	25.87	2631.7	72.93
		Permuted	63.93	64.93	1.00	35.07	29.50	2811.7	69.60
	8k	Base	80.40	83.10	2.70	16.90	14.30	3837.4	85.20
		Aligned	82.47	86.67	4.20	13.33	11.83	3449.6	86.73
		Permuted	80.77	83.80	3.03	16.20	14.27	3802.4	85.60
	16k	Base	88.60	94.10	5.50	5.90	5.10	4626.5	92.00
		Aligned	88.60	96.27	7.67	3.73	3.43	4063.0	92.53
		Permuted	88.73	94.73	6.00	5.27	4.63	4574.7	92.40
	32k	Base	90.10	99.40	9.30	0.60	0.60	4988.9	93.40
		Aligned	89.57	99.83	10.27	0.17	0.17	4286.2	93.20
		Permuted	90.50	99.50	9.00	0.50	0.40	4900.5	93.67
4B	4k	Base	62.80	63.10	0.30	36.90	28.60	2862.4	67.40
		Aligned	68.30	68.53	0.23	31.47	25.90	2682.5	72.07
		Permuted	63.50	63.60	0.10	36.40	30.13	2854.3	67.80
	8k	Base	85.30	86.60	1.30	13.40	11.80	3772.5	89.20
		Aligned	87.70	88.80	1.10	11.20	9.30	3478.7	90.87
		Permuted	85.50	86.20	0.70	13.80	11.30	3796.6	89.27
	16k	Base	93.80	96.00	2.20	4.00	3.80	4377.3	96.20
		Aligned	94.27	96.47	2.20	3.53	2.97	3977.3	96.60
		Permuted	93.93	95.77	1.83	4.23	3.80	4401.0	96.33
	32k	Base	96.00	99.40	3.40	0.60	0.40	4659.1	97.60
		Aligned	96.10	99.80	3.70	0.20	0.17	4188.6	97.80
		Permuted	96.13	99.57	3.43	0.43	0.33	4688.0	97.87

Table 29: Additional paired analyses of MATH-500 (K2). Token-use and budget-interaction intervals were computed after the prespecified accuracy analyses. The budget interaction is the 4k pairing contrast minus its 32k counterpart, evaluated directly on paired outcomes.

Model	Quantity	Difference	95% interval
Qwen3-1.7B	4k–32k change in $A - P$ (pp)	+5.23	$[+3.43, +7.10]$
	Mean tokens, $A - P$	-614.24	$[-825.62, -421.40]$
	Mean tokens, $A - B$	-702.71	$[-937.15, -477.09]$
Qwen3-4B	4k–32k change in $A - P$ (pp)	+4.83	$[+3.07, +6.73]$
	Mean tokens, $A - P$	-499.35	$[-683.09, -309.20]$
	Mean tokens, $A - B$	-470.49	$[-632.30, -309.64]$

E TARGET GEOMETRY, NUMERICAL VALIDATION, AND STATISTICAL INFERENCE

This appendix supports the formal construction in Section 3, the controlled interventions in Section 5, and the uncertainty summaries throughout the paper. The derivations establish the relation between centered displacements and targets; numerical checks verify the realized intervention; paired inference specifies the population and units behind each result.

E.1 WHAT CENTERING AND DIVERGENCE MATCHING PRESERVE

For positive full-vocabulary distributions, adding a constant to a logit displacement leaves its normalized target unchanged. Centering under q_0 in Equation 3 therefore fixes a representation of the same teacher behavior. Since $d_C = \log(q_C/q_0) - c(s)$, we have $q_0 \exp d_C = e^{-c(s)} q_C$; normalization yields Equation 4 exactly. Recentring a donor under the recipient reference likewise leaves its softmax target invariant but defines its norm in the recipient distribution.

The local expansion explains the numerical initialization. For a centered direction d with $\mathbb{E}_{q_0} d = 0$, define $q_\alpha(a) = q_0(a)e^{\alpha d(a)}/Z(\alpha)$, where $Z(\alpha) = \sum_{v\in\mathcal{V}} q_0(v)e^{\alpha d(v)}$. Then

$$q_\alpha(a) = q_0(a)\big(1 + \alpha d(a)\big) + O(\alpha^2), \qquad J(q_0, q_\alpha) = \frac{\alpha^2}{8}\mathbb{E}_{q_0}[d(a)^2] + O(\alpha^3). \tag{15}$$

The first identity follows by differentiating the normalizer at zero; the second is the second-order expansion of the two KL terms about their common distribution. The norm $\sqrt{\mathbb{E}_{q_0} d^2}$ thus gives a local scale estimate. This local expansion supplies the initialization; the final target is checked using its actual finite-displacement JSD. The identities provide the algebraic basis of the experimental construction.

Reference-distance matching constrains the target displacement while allowing its optimization consequences to vary. For student logits z with $p = \mathrm{softmax}(z)$ and detached target q, direct differentiation gives

$$\frac{\partial J(p,q)}{\partial z_a} = \frac{1}{2}p(a)\left[\log\frac{p(a)}{m(a)} - \mathbb{E}_{v\sim p}\log\frac{p(v)}{m(v)}\right], \qquad m = \tfrac{1}{2}(p+q). \tag{16}$$

Two targets at the same reference-relative divergence can have different logit gradients, target entropies, and parameter gradients. The parameter Jacobian adds another dependence on state. The controlled quantity is the reference-relative distributional distance; differences in the resulting optimization dynamics belong to the effect of the target-assignment intervention.

E.2 ELIGIBILITY AND NUMERICAL SAFEGUARDS

Each fixed pair is evaluated in both directions at the same response index. Candidate positions must exist in both generated responses and share the same *pre-token* phase. The phase is determined by tokens strictly before the candidate; predicting `</think>` itself remains a thinking-phase prediction. Semantic steps may differ across paired positions, within the matched construction family and generation phase.

Target construction uses detached teacher arithmetic promoted to float32 before log-softmax. The donor scale is initialized from the recipient-reference norm and refined using two multiplicative corrections followed, where needed, by safeguarded bracketing and bisection of the actual JSD discrepancy. The implementation permits bracket growth by a factor of four, at most eight bracket expansions, at most twelve bisection steps, and a maximum numerical scale of 1,024. Scale-tail checks use threshold 32 and a 1% tail allowance. The 4B run implements this allowance through exact pooled counts, as detailed below.

A position is routed only if its constructed finite-vocabulary target satisfies the relative-plus-absolute tolerance in Equation 9. Matched neutral positions are distinguished from matched nonzero signals. Degenerate directions, an unbracketed target, or an exhausted solve do not silently fall back to a different teacher: the position is removed from the common candidate loss support. The same-phase fraction must be at least 0.50 of the valid length intersection, and the matched common-support

fraction at least 0.95 of the same-phase candidates. Receipts additionally check finite losses and gradients and nonzero adapter updates.

Equation 2 is normalized by the eligible-position count within each rank-local loss batch before distributed gradient aggregation. The pooled token counts used for numerical checks are accumulated separately from this loss normalization.

The shared controls operate at two levels. For a given sampled batch, both candidate target constructions are evaluated on the same accepted positions. Across separate training runs, the students generate their own trajectories, so their realized positions, donor states, and cumulative token counts need not coincide. At *evaluation*, all continuation policies receive the same frozen Base histories. The latter is an exact history intervention; the former is a common construction rule inside an on-policy training intervention.

E.3 Exact specification of the two shaping controls

For SelfSharp, the calibration solves a scalar, reward-blind target-distance problem on 16 fixed training examples before optimization:

$$\mathbb{E}_{s\in\mathcal{S}_{\mathrm{cal}}} J(q_0, q_{\mathrm{sharp},\tau}) \;\simeq\; \frac{1}{3}\sum_{c\in\{T,O,F\}} \mathbb{E}_{s\in\mathcal{S}_{\mathrm{cal}}} J(q_0, q_c). \tag{17}$$

Here $\mathcal{S}_{\mathrm{cal}}$ is the shared set of calibration states and T, O, F index Target, Other, and Fixed contexts. A float32 bisection search fixes $\tau < 1$: the calibrated temperatures are 0.4482181668 for Qwen3-1.7B and 0.5255885720 for Qwen3-4B. Each remains constant during training. Only the target logits are temperature-scaled; the student is not softened, and there is no conventional temperature-squared loss multiplier. This is aggregate initialization matching, unlike the per-recipient actual-divergence matching of StatePermuted.

To make the comparison in Section 4.3 explicit, the negative-advantage control implements the entropy-adaptive, low-probability-token update of Ding & Zhang (2026). The main-text name Tail-Suppress describes this published operation. At both scales, each data-parallel loss call contains one response; four such calls are accumulated before an optimizer step. Selection and entropy normalization are therefore performed within that response. Let $\mathcal{I}$ contain the $\max(1, \lfloor 0.2n \rfloor)$ lowest current-actor sampled-token log-probabilities among the n valid tokens in a data-parallel local batch. Ties are resolved by stable row-major ordering. For selected positions, write $H_t = -\sum_a p_\theta(a \mid s_t) \log p_\theta(a \mid s_t)$ and let $H_{\min}, H_{\max}$ be the extrema *within the selected set*. The detached advantage is

$$A_t = -\frac{1}{2} - \frac{1}{2}\frac{H_t - H_{\min}}{H_{\max} - H_{\min}}, \qquad \mathcal{L}_{\mathrm{tail}} = -\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \mathrm{sg}[A_t] \log p_\theta(y_t \mid s_t). \tag{18}$$

When the selected entropy range is at most 10^{-12}, all selected advantages are -1, following the implementation convention. Unselected tokens enter neither numerator nor denominator. The control uses the student distribution alone. Its recipe is fixed as a secondary comparison before evaluating the contextual-policy gap.

E.4 Paired inference and endpoint accounting

For each contrast, decode histories are first averaged within each problem and training seed. A bootstrap replicate samples problem indices with replacement and independently samples the registered training-seed indices with replacement, then evaluates the contrast over their Cartesian product. Conditions remain paired on the same problems and seeds; the Base outcome is reused, not resampled as if it had independent training runs. The shared-prefix and source-transfer tables use a standardized 20,000-replicate implementation with NumPy default_rng(0), 95% equal-tail percentile intervals, and the finite three-training-seed set. Tables that retain a prespecified 10,000/90% or 20,000/95% analysis identify that specification in their captions. The standardized 20,000/95% analysis uses the same endpoint observations; changing the resampling implementation can move interval endpoints while preserving the point estimate. Multiple analysis versions of one comparison constitute one experiment.

The intervals summarize paired problem variation and the observed finite set of three training seeds. Seed-specific effects accompany the pooled estimate to expose that finite training variation. Ab-

sorbed histories remain in the average and contribute zero to policy contrasts. Each comparison reports its number of distinct questions; policies, temporal windows, and repeated history records preserve their dependence structure.

For complete-response samples with c_i correct out of n draws per problem i, the supplementary Pass@k estimator (Chen et al., 2021) is

$$\widehat{\mathrm{Pass@}k} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}}\right], \qquad 1 \le k \le n. \tag{19}$$

We use $\binom{n-c_i}{k} = 0$ when $n - c_i < k$. Pass@k measures success in a finite group of sampled attempts; Avg@K measures per-sample correctness under the declared generation or continuation protocol.

Boundary contrasts compare the four complete continuation *regimes*, not only histories that cross the boundary. Their decomposition is algebraic at the regime-value level. A pre-policy effect includes changed timing and changed prefixes delivered to the post-policy. Conditioning on observed completion, boundary crossing, or a correct intermediate answer would instead select post-treatment subsets and answer a different question.

E.5 Qwen3-4B training diagnostics and scale-tail aggregation

All six 4B training cells complete 100 optimizer steps with finite losses, finite positive gradient norms, and nonzero LoRA updates. The same-phase fraction of length-intersected positions is 99.69–99.93%, and approximately 99.985% of same-phase positions satisfy the actual-JSD matching rule. Table 30 gives the pooled fraction of accepted positions whose donor scale exceeds 32.

Table 30: Qwen3-4B scale-tail diagnostics. Each entry is the percentage of common-support scale observations above 32, pooled over all distributed ranks and loss evaluations in a training cell. The acceptance threshold is 1%.

Training arm	Seed 51	Seed 52	Seed 53
Aligned	0.5586	0.5565	0.5579
StatePermuted	0.5768	0.5839	0.5772

The initial 4B execution stopped after its first training cell because its scale-tail check required every rank-local loss evaluation to have a 99th-percentile scale at most 32. Evaluation had not begun. Before the corrective run, the aggregation was changed to the fraction of all accepted scale observations exceeding 32:

$$f_{>32} = \frac{\sum_{r,c} n_{>32}^{(r,c)}}{\sum_{r,c} n_{\mathrm{scale}}^{(r,c)}} \le 0.01, \tag{20}$$

where r indexes distributed ranks and c indexes loss evaluations. The additive numerator and denominator make this criterion invariant to how the same observation population is partitioned. Threshold 32 and the 1% tail allowance were retained. Model initialization, data, seeds, target construction, matching tolerances, optimization, and endpoint analysis were also retained. All six reported cells were trained afresh from Base, and the initial cell remains outside the reported result. This appendix records the aggregation change separately from the state-permutation intervention.

E.6 Same-question matching and rank-local zero-signal handling

All six 1.7B same-question training cells complete 100 optimizer steps, with finite gradients and nonzero adapter updates bound to their evaluation checkpoint identities. Among eligible same-phase, prefix-diverged candidates, pooled matching support is 99.951–99.956%; the fraction with scale above 32 is 0.137–0.140%. These are conditional matching diagnostics; the prefix-divergence and phase filters define the candidate set first.

Of 4,632 rank-level loss calls, 19 have no eligible same-question signal (0.410%). They contribute a graph-connected zero loss locally while other ranks retain their eligible contributions. Metric-name aggregation takes the cross-rank union, preserving pooled count-based validity checks. This

convention preserves the distributed computation and the pooled diagnostic denominator when one rank has no eligible positions.

The audit separately validates 9,216 paired seed-history records for the Aligned/Base crossovers: pre-phase hashes, lengths, and boundary states match within the fixed-pre-policy comparisons. The conditional post-call results use the corrected seed-aware selector described in Appendix D.1.

E.7 Matching support across pairing scopes and objectives

The same-question 4B and cross-question reverse-KL comparisons each contain six fresh step-100 policies. All twelve training receipts bind finite losses, finite positive gradient norms, and nonzero adapter updates to their evaluated checkpoints. Table 31 reports realized ranges over the six training cells per comparison. Ratios use the stated candidate-set denominators.

Table 31: Realized matching and scale-tail diagnostics (percent). Ranges retain the precision of the dedicated experiment reports.

Diagnostic	Same-question 4B	Reverse KL, 1.7B
Same phase / length intersection	99.64–99.89	99.5486–99.9406
Diverged prefix / same phase	99.12–99.18	Not required
Matched / eligible candidates	99.9855–99.9879	99.9476–99.9518
Pooled scale above 32	0.476–0.494	0.1482–0.1573

For same-question matching, eligible candidates have both same-phase and diverged prefixes; for reverse-KL cross-question matching they have same-phase prefixes. Both retain the 95% matching-support threshold and 1% pooled-tail threshold. The same-question cells record 772 rank-level loss calls each and no zero-signal microbatches in their final cumulative records. Reverse-KL cells record 744 such calls each. Epoch-tail accumulation explains these counts while every cell reaches exactly 100 optimizer steps. Local minibatch scale quantiles remain descriptive under the pooled-token tail rule in Equation 20.

E.8 Analysis specifications and interval sensitivity

Table 32 retains the shaping study's prespecified interval calculation beside the standardized reporting rule above. Its SelfSharp–Base interval lies inside the prespecified ± 2-point operational-equivalence band, which characterizes this particular training recipe and continuation readout. The negative-advantage comparison is secondary. The state-pairing intervention in Section 5 supplies the direct test of state assignment. The source-transfer intervals in Table 8, parameter-recomposition

Table 32: Shaping contrasts under the prespecified 90% analysis (percentage points). These are the same 1.7B endpoint observations as Table 6, which reports the standardized 95% analysis. ContextMean averages policy outcomes; the parameter centroid is defined in Equation 11.

Contrast	Effect	90% interval
ContextMean − Base	3.84	$[2.17, 5.62]$
SelfSharp − Base	-0.26	$[-1.95, 1.43]$
TailSuppress − Base	-0.72	$[-2.60, 1.11]$
ContextMean − SelfSharp	4.10	$[2.52, 5.79]$
ContextMean − TailSuppress	4.56	$[3.10, 6.10]$

intervals in Table 9, and temporal intervals in Appendix D.3 likewise retain their stated analysis specifications. Comparing these entries requires their cohort, metric, budget, and coverage, rather than treating a change in coverage as a new empirical observation.

Seed-aware secondary selection. For the actual-answer-policy-call analysis in Appendix D.1, selection uses the full seed–problem–history key and distinguishes post-policy invocation from a boundary already present in an absorbed history. A cross-seed selection used during analysis had included 277 and 124 non-calling seed-history slots in the two AA/AB comparisons. The reported

secondary estimates use the corrected selector; the full-cohort primary estimates and pre-phase pairing are unchanged. This affects the definition of a selected secondary population, not the endpoint sample or the trained policies.

F Connections to Existing Distillation Mechanisms

This section expands Section 2 around the three distinctions tested in the main text: the source of reference context, the form of token supervision, and the state to which that supervision is assigned.

From rationale imitation to learner-state supervision. Sequence-level distillation trains on teacher-generated outputs (Kim & Rush, 2016). Reasoning distillation extends this approach to chains of thought (Magister et al., 2023) and uses teacher rationales as additional multitask supervision (Hsieh et al., 2023). OPD instead queries teacher predictions at learner-generated prefixes (Agarwal et al., 2024; Gu et al., 2024); DistiLLM combines skew-KL objectives with adaptive off-policy reuse of student outputs (Ko et al., 2024). This progression separates the source of training trajectories from the loss used to align predictions, providing context for our controlled state-assignment and objective comparisons.

Privileged information and self-generated references. Learning with privileged information uses training-only explanations to support a student that acts without them (Vapnik & Izmailov, 2015); generalized distillation connects this setting to teacher–student knowledge transfer (Lopez-Paz et al., 2016). Context distillation instantiates this idea for language models (Snell et al., 2022), while OPSD and SDPO query context-enriched self-teachers on student-generated trajectories (Zhao et al., 2026a; Hübotter et al., 2026). References can also be constructed from the learner's own outputs: self-consistency aggregates sampled reasoning paths (Wang et al., 2023), and U-OPSD uses a consensus-derived pseudo-solution to supervise disagreeing completions (Li et al., 2026d). These approaches vary the source of additional information; our source comparisons examine how that choice changes continuation gains.

Reusable context and training-state coverage. Ichihara et al. (2026) establish that worked solutions from other problems can remain useful in self-distillation. Section 4.2 studies this reference-reuse phenomenon through source-specific continuation comparisons and the transfer tests in Appendix B. One- and few-query studies (Fu et al., 2026; Hou et al., 2026) relate small training sets to state exposure, long reasoning, and the rate of alignment; data-free distillation also examines reduced dependence on externally supplied query content (Li et al., 2026b). These results motivate separating the reference supplied to a teacher from the feedback it produces at each visited prefix.

Compatibility and the reliability of token feedback. Simple-OPD connects effective warm-up to teacher-compatible reasoning patterns (Liu et al., 2026). Analyses of student–teacher mismatch and privileged bias describe cases in which a distillation signal diverges from task correctness (Wang et al., 2026a; Harne et al., 2026); token-credit diagnostics examine this relationship directly (Nguyen et al., 2026). TGOPD verifies teacher reliability at the prompt level (Zhang et al., 2026), VISTA adapts the teacher using verified student rollouts (Ding et al., 2026), and FutureBridge-OPD evaluates guidance through subsequent agent trajectories (Chen et al., 2026). RP-OPSD instead concentrates supervision around reasoning pivots identified from matched teacher views (Wang et al., 2026b). These studies connect useful feedback to the learner's reasoning context. Our same-question reassignment holds the problem and reference fixed to test dependence on the particular partial solution.

Probability shaping and reference-relative changes. Ding & Zhang (2026) find that suppressing low-probability sampled tokens with negative advantages can reproduce substantial OPD gains, motivating an entropy-adaptive self-adaptation rule. Section 4.3 evaluates this operation as TailSuppress under the exact-prefix reasoning protocol, with its local-batch implementation in Appendix E.3. The direct contrasts in Appendix B.1 specify how that rule compares with contextual supervision in these experiments. CompassOPD (Gu et al., 2026) separates within-family shifts from cross-family offsets, and RouteOPD (Xia et al., 2026) studies probability transport between excess and deficient token mass. These approaches support analyzing a teacher-induced change separately from an absolute teacher distribution.

TV-OPD studies advantage signs and bounded update signals (Xiao et al., 2026); Negative Self-Distillation constructs a flawed self-teacher and selectively trains away from its predictions (Pei et al., 2026). OPD^2 uses the probability gap between post-trained and base teachers (Heo et al., 2026), while weak-to-strong OPD constructs a proxy teacher from contrasting smaller models (Yu et al., 2026a). OPRD uses the teacher's policy shift to amplify verifier-supported student updates (Park et al., 2026). Together with Direct-OPD, these methods motivate treating relative changes in predictions as transferable supervisory objects. Our matched-divergence test examines the state dependence of a context-induced change.

Reasoning trajectories and finite computation. Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022), while DeepSeekMath and DeepSeek-R1 develop reasoning through outcome-based reinforcement learning (Shao et al., 2024; Guo et al., 2025). Work combining OPD and RLVR studies the sequencing of dense and outcome-based supervision (Li et al., 2026a). Complementary analyses examine how the organization and completeness of reasoning traces affect learning: Aha-Flow varies discourse markers and the organization of privileged supervision (Wang & Peng, 2026), and Hwang et al. (2026) study the utility of partial traces and trajectory endpoints. At inference, compute allocation and budget forcing explicitly change the amount or organization of reasoning (Snell et al., 2025; Muennighoff et al., 2025); L1 learns to satisfy requested length constraints through reinforcement learning (Aggarwal & Welleck, 2025). Harmful-overthinking analyses examine failures after a correct answer has been reached (Caldarella et al., 2026). These connections motivate reporting both terminal accuracy and completion behavior across token budgets; our policy crossovers locate where the learned pairing advantage is expressed during generation.

From a guidance signal to its assignment. Our intervention uses the reference-relative change as an experimental object. It preserves the recipient prediction, matches the finite reference divergence, and varies the correspondence between the change and its receiving state. The same-question experiment fixes both problem and reference, while reverse-KL training changes the optimization objective under the same target calibration. The policy crossovers then connect these training interventions to the deployed reasoning phase. The resulting contribution is a controlled account of the functional guidance–state relation, complementary to studies of source selection, probability shaping, and transferable policy shifts.