BEYOND THE ANSWER: STUDENT-CONDITIONED TRACE CONSTRUCTION FOR REASONING DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A reasoning trace can accompany a correct answer without providing effective supervision for a particular student. We study how to construct traces that improve learning beyond answer-only supervision by considering both their intermediate content and their accessibility to the target learner. We introduce REDIST, a student-conditioned trace-construction and distillation framework for evidence-rich reasoning. Starting from the student's own attempts, REDIST revises answer-correct drafts against the supplied evidence and repairs unsuccessful drafts with additional guidance from the verified answer. A frozen student-matched reader screens and ranks original and edited candidates using answer recovery and verified-answer probability, given the question, evidence, and candidate trace. We distill the selected targets while balancing trace and final-answer supervision with an auxiliary answer loss. Construction and selection occur offline, and inference uses only the distilled student. Experiments on question-answering benchmarks show that REDIST can improve both source-task and cross-benchmark performance over answer-only, original-trace, and the evaluated trace-selection and rewriting baselines. Across supervision recipes, performance rankings can differ between source and transfer benchmarks, highlighting the importance of evaluating trace teaching utility across distributions.

1 INTRODUCTION

Reasoning-trace distillation trains a language model on intermediate derivations as well as final answers, making generated solutions a practical source of supervision (Hsieh et al., 2023; Li et al., 2025). In evidence-rich tasks such as multi-hop question answering, a derivation can communicate supporting facts and intermediate relations that are useful on new questions. Yet the trace also becomes a sequence the student must learn to predict: missing reasoning steps, unnecessary elaboration, or unfamiliar expression can make it an ineffective teaching target even when its final answer is correct. The practical challenge is therefore *to construct reasoning traces that improve the student's learning beyond answer-only supervision.*

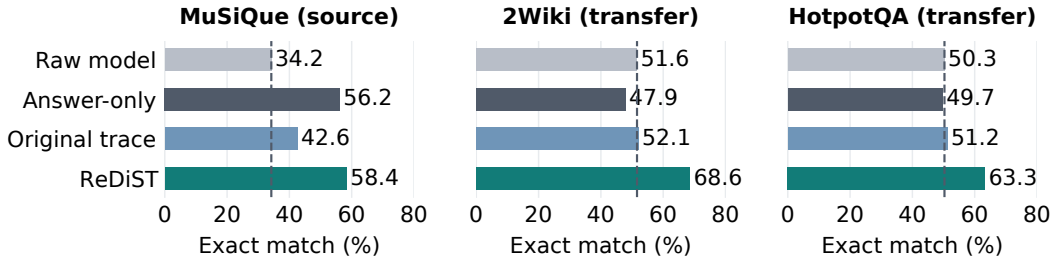


Figure 1: **Source-task accuracy and cross-benchmark transfer respond differently to supervision.** Qwen3-8B before task-specific training (Raw) and after training on MuSiQue with answer-only supervision, original student traces, or REDIST. MuSiQue measures source-task performance, while 2Wiki and HotpotQA measure cross-benchmark performance. Dashed lines mark each benchmark's raw-model EM. Values are EM point estimates from Table 1. Comparisons between trained variants reflect complete training recipes.

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

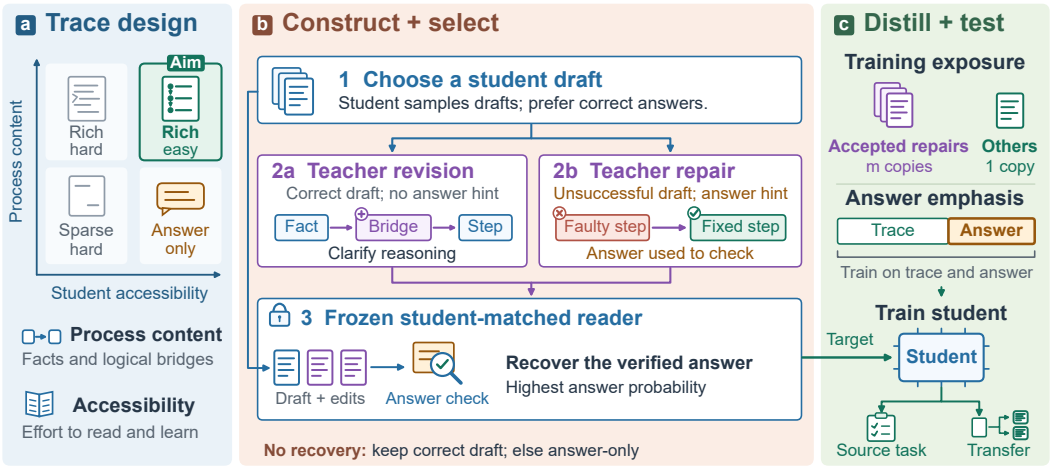


Figure 2: **REDIST: student-conditioned trace construction and distillation.** (a) Process content and student accessibility guide the design. (b) A rewrite model revises answer-correct student drafts against the supplied evidence and repairs unsuccessful drafts with additional guidance from the verified answer. A frozen student-matched reader screens and ranks original and edited candidates using answer recovery and verified-answer probability, given the question, evidence, and candidate trace. (c) Repair exposure m and auxiliary answer-loss weight α control how selected targets enter training. Teaching utility is evaluated through the trained student's source-task and cross-benchmark performance.

Figure 1 makes this challenge concrete. With MuSiQue supervision, answer-only training raises Qwen3-8B's source-task exact match (EM) from 34.2% to 56.2%, while its EM on 2WikiMultiHopQA (2Wiki) falls from 51.6% to 47.9%. Training on the original student traces yields a different profile: 42.6% source EM and 52.1% transfer EM. The answer-only and original-trace recipes therefore rank differently on the two benchmarks: *higher source-task accuracy does not necessarily imply better cross-benchmark performance.* These contrasting profiles motivate assessing trace supervision through both source-task and cross-benchmark performance.

Prior work improves reasoning supervision through self-training and correction (Zelikman et al., 2022; Dou et al., 2024; Kapusuzoglu et al., 2026), including metacognitive rewriting of student attempts (Kim et al., 2026). Other methods select trajectories for their informativeness and learnability (Yang et al., 2026; Yu et al., 2026), preserve the student's style during refinement (Qian et al., 2026), or use student answer prediction to optimize explanations and select rationales (Cetin et al., 2025; Patnaik et al., 2025). These studies establish the importance of both trace content and compatibility with the learner. Building on these insights, we focus on a specific construction problem in evidence-rich reasoning: *deciding which evidence-guided revisions and repairs of the target student's own attempts should become its teaching targets.* We couple evidence-guided editing with selection based on the frozen student's response, and examine what this construction pipeline adds beyond selecting unedited student drafts while accounting for the allocation of trace and answer supervision.

These observations motivate the following question:

> *How can we turn the target student's own attempts into reasoning supervision that improves learning beyond final-answer labels?*

We approach this question through two considerations. **Process content** concerns the supporting facts and intermediate relations communicated beyond the final-answer label, while **student accessibility** concerns the target student's ability to use and learn from that content. A conclusion-only trace can reveal the answer while omitting the reasoning that supports it, whereas a detailed derivation can supply that reasoning in a form the student struggles to learn. Improving a derivation's content therefore need not make it a better teaching target for a particular student.

To address these considerations, we introduce REDIST (*Refinement and Distillation of Student Traces*), a student-conditioned trace-construction and distillation framework for evidence-rich reasoning (Figure 2). *Its organizing principle is to let the target student shape both the starting drafts and*

the selection of teaching targets. A rewrite model revises answer-correct drafts using the question, supplied evidence, and original attempt. For unsuccessful drafts, it additionally receives the verified answer as guidance for repairing the supporting derivation against the evidence. The two branches address complementary needs: *improving intermediate supervision even when the student's final answer is correct, and recovering supervision for questions it initially fails to solve.*

A frozen student-matched reader then receives the question, evidence, and each original or edited candidate trace, and screens and ranks candidates using answer recovery and verified-answer score. This supplies a practical selection signal, while teaching utility is evaluated through the student's performance after distillation. During training, repair exposure controls how often accepted repairs are seen, while an auxiliary answer loss controls the emphasis on final-answer prediction alongside trace supervision. These choices separate which targets are constructed from how they contribute to optimization. Construction and selection occur offline, and inference uses only the distilled student.

Under MuSiQue supervision, REDIST improves source-task and 2Wiki performance over answer-only and original-trace supervision across four students from two model families (Table 3). In the primary Qwen3-8B comparison, it achieves the highest reported point estimates across all four benchmarks, including 58.4% MuSiQue EM and 68.6% 2Wiki EM (Table 1). A separate Qwen3-8B comparison extends these gains to a 2Wiki supervision and transfer to MuSiQue (Table 7). Component comparisons show additional gains beyond selecting unedited student drafts and highlight the importance of answer-loss allocation (Table 2).

Our contributions are threefold:

- **Formulation and evaluation principle.** We formulate incremental trace teaching utility relative to matched answer-only supervision for a specified student and evaluation distribution, distinguishing this quantity from complete-recipe performance comparisons.
- **Student-conditioned trace construction.** We introduce REDIST, which combines evidence-guided revision and repair of student drafts with selection by a frozen student-matched reader and explicit control of repair exposure and answer supervision.
- **Empirical evidence.** We demonstrate gains over answer-only and original-trace supervision across four target students and two model families. Component comparisons examine how trace construction and answer-loss allocation affect source-task and transfer performance.

2 PROBLEM SETTING AND TRACE TEACHING UTILITY

2.1 EVIDENCE-RICH REASONING DISTILLATION

We consider reasoning tasks with training data $\mathcal{D} = \{(Q_i, E_i, Y_i)\}_{i=1}^{n}$, where Q_i is a question, E_i is supplied evidence, and Y_i is a verified answer. The evidence may contain both supporting information and distractors. Given an initial student s_0, our goal is to *construct reasoning supervision that improves its learning beyond final-answer labels alone.* Construction uses these training examples and generated attempts, without requiring gold reasoning traces. At inference, the trained student answers new questions from supplied evidence.

We distinguish a generated attempt $(R, \widehat{Y})$ from a supervision target $(R^\star, Y)$, where R is the generated reasoning, $\widehat{Y}$ is the generated final answer, and $R^\star$ is the constructed trace paired with the verified answer Y. On a fixed training cohort, a construction c specifies the targets $R_i^\star$, allowing an empty trace for answer-only supervision. A recipe π specifies example exposure, loss allocation, and the update budget. This separates which traces are used from how they enter training.

2.2 STUDENT-RELATIVE TEACHING UTILITY

Let $M_{\mathcal{P}}(s)$ denote a higher-is-better performance metric for student s on evaluation distribution $\mathcal{P}$ under a fixed evaluation protocol. Write $\mathrm{Train}_{s_0}(c; \pi)$ for training from s_0 with construction c and recipe π. For a given c, its paired answer-only construction $c_\varnothing$ removes each trace while retaining the same examples and verified answers. We define incremental trace teaching utility as

$$\mathrm{UTIL}^{M}_{s_0,\mathcal{P}}(c; \pi) = M_{\mathcal{P}}(\mathrm{Train}_{s_0}(c; \pi)) - M_{\mathcal{P}}(\mathrm{Train}_{s_0}(c_\varnothing; \pi)) \,. \tag{1}$$

The paired runs retain the same question IDs, evidence, answer labels, initialization, optimizer settings, update budget, and realized per-example sampling schedule, as well as the same input and output templates apart from removing the trace content. In particular, removing a trace does not change its example's exposure. The same loss rule applies to the resulting completions, although their supervised-token counts can differ, and matching token budgets is a separate control. This quantity measures the effect of including the constructed traces under the specified training protocol. Complete-recipe comparisons may also change the cohort, exposure, or loss allocation and are interpreted separately. Evaluating Equation (1) on the source distribution and on other benchmarks yields source-task and cross-benchmark utility, respectively. A construction can therefore have different teaching utility across evaluation distributions.

2.3 Design Considerations: Content and Accessibility

Process content. Process content concerns the supporting facts and intermediate relations expressed in a trace beyond the final-answer label. These elements may already appear in the supplied evidence, and the trace can make their relevance and connections explicit. A conclusion-only trace omits those connections, while a more elaborate trace can connect a factual relation to the wrong intermediate entity. Assessing content therefore requires examining both the supporting statements and their connection to the question. Figure 3 illustrates how a trace can identify relevant entities while introducing an unsupported relation.

Student accessibility. Accessibility concerns how readily the target student can use and learn from the expressed content. Wording, organization, and the granularity of intermediate steps can affect this compatibility. Familiar expression can preserve an incorrect step, whereas a valid derivation can remain difficult for a particular student to learn.

Content and accessibility are distinct considerations whose effects can interact, and neither alone determines downstream teaching utility. They therefore need to be considered together when constructing teaching targets. This motivates the framework in Section 3, which combines evidence-guided editing of the supporting derivation with candidate selection informed by the target student's response.

3 The ReDiST Framework

ReDiST treats trace distillation as a student-specific teaching problem: the target student shapes both the starting drafts and the selection of teaching targets. The framework has three stages, illustrated in Figure 2: (1) evidence-guided revision or answer-assisted repair of student drafts, (2) screening and ranking by a frozen student-matched reader, and (3) distillation, where repair multiplicity m controls example exposure and answer-loss weight α controls answer emphasis. Algorithm 1 in Appendix B.2 summarizes this offline training pipeline. At inference, only the distilled student is used.

3.1 Student-Draft Revision and Repair

Draft selection. For each training example (Q, E, Y), the initial student s_0 generates K_d attempts $(R^{(k)}, \widehat{Y}^{(k)})$, $k = 1, \ldots, K_d$. We use a frozen student-matched reader to assign each eligible draft a verified-answer score $q_{s_0}(Y \mid Q, E, R)$, defined in Section 3.2. Among complete, eligible drafts, we first consider those whose generated final answer matches Y and choose the one with the largest score. If none answers correctly, we choose the highest-scoring eligible draft. Denote the selected draft by $(R^{(0)}, \widehat{Y}^{(0)})$. Cases with no reader-eligible draft, along with other parsing exceptions and format-only exclusions, follow the deterministic rules in Appendix B.2.

Evidence-guided revision. When $\widehat{Y}^{(0)}$ matches Y, the rewrite model (teacher) g receives $(Q, E, R^{(0)})$ and generates K_r candidate revisions. It is instructed to retain supported reasoning, remove irrelevant or unsupported content, and make intermediate entities and relations explicit. The verified answer is not supplied as a separate field in this branch. The teacher may change individual steps, their ordering, and their expression to improve the derivation.

Answer-assisted repair. When the answer of the selected draft does not match Y, the teacher additionally receives Y as a repair target and generates K_r candidates. It is instructed to correct the supporting derivation against E and connect the relevant facts to the verified answer. This branch

makes unsuccessful attempts available for trace supervision beyond correctness-only filtering. Invalid edited outputs are removed by the deterministic rules in Appendix B.2, and each candidate retains its origin as an original draft, a revision, or a repair.

3.2 Frozen-Student Target Selection

An edit can change both the derivation and the student's response to it. We therefore compare edited candidates with the selected draft before assigning a teaching target. The reader remains frozen throughout construction and distillation. Given (Q, E, R), it produces an answer prediction $\widehat{Y}^{\mathrm{read}}_{s_0}(Q, E, R)$ and a verified-answer score $q_{s_0}(Y \mid Q, E, R)$. We screen candidates by answer recovery and rank those that pass by their verified-answer score. The task-specific reader interface and scoring rules are given in Appendix B.2.

Let $\mathcal{C}$ contain the selected draft and edited candidates that satisfy the validity rules for reader input, with each candidate's origin retained. Define the candidates that pass the recovery check as

$$\mathcal{C}^+ = \left\{ R \in \mathcal{C} : \widehat{Y}^{\mathrm{read}}_{s_0}(Q, E, R) \text{ matches } Y \right\}. \tag{2}$$

We select the supervision trace according to

$$R^\star = \begin{cases} \arg\max_{R \in \mathcal{C}^+} q_{s_0}(Y \mid Q, E, R), & \mathcal{C}^+ \neq \varnothing, \\ R^{(0)}, & \mathcal{C}^+ = \varnothing \text{ and } \widehat{Y}^{(0)} \text{ matches } Y, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{3}$$

Candidates retain their identities and origins even when their trace texts coincide, while score ties follow a fixed candidate order. If no candidate passes, we retain an initially answer-correct draft or use an empty trace for answer-only supervision. Every retained example remains paired with the verified final answer Y.

The reader's recovery provides a student-conditioned selection signal. Because the reader also receives the question and evidence, successful recovery does not by itself establish stepwise correctness or learning benefit.

3.3 Distillation and Supervision Allocation

After selecting teaching targets, we control how often accepted repairs enter training and how strongly final-answer tokens are supervised.

Repair exposure. Each accepted repair appears m times in the training mixture, and every other training example, including answer-only fallbacks, appears once. During training, we sample mini-batches from the training mixture.

Answer supervision. Each supervised completion contains the selected trace $R_i^\star$ followed by the verified answer Y_i. Under a token-averaged completion loss, a long trace contributes many more supervised positions than a short answer. To make the emphasis on final-answer prediction explicit, we combine the full-completion loss with an auxiliary answer-token loss.

For a training batch, let $\mathcal{T}$ contain all supervised completion positions and let $\mathcal{A} \subseteq \mathcal{T}$ contain the answer-content positions. For trainable student parameters θ, let $\ell_t(\theta)$ denote the next-token cross-entropy at position t, conditioned on the question, evidence, and preceding completion tokens. We define $\mathcal{L}^{\mathrm{full}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \ell_t(\theta)$ and $\mathcal{L}^{\mathrm{ans}} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \ell_t(\theta)$. The distillation objective is

$$\mathcal{L} = (1 - \alpha)\mathcal{L}^{\mathrm{full}} + \alpha \mathcal{L}^{\mathrm{ans}}, \qquad 0 \leq \alpha \leq 1. \tag{4}$$

The full-completion loss already includes answer tokens, so α controls their additional emphasis. At $\alpha = 1$, only answer positions receive direct token supervision, but their predictions still condition on the preceding trace. This therefore differs from the trace-removed answer-only control in Equation (1).

Connection to content and accessibility. Draft-conditioned editing and frozen-student selection jointly address the content and accessibility considerations in Section 2. Repair multiplicity m and answer-loss weight α control how the selected supervision enters training.

Table 1: **Source-task and cross-benchmark performance after MuSiQue supervision.** All systems use Qwen3-8B. Trained variants use only MuSiQue examples for gradient updates.

	MuSiQue (Source)		2Wiki (Transfer)		HotpotQA (Transfer)		QuALITY (Transfer)	
Method	**EM↑**	**F1↑**	**EM↑**	**F1↑**	**EM↑**	**F1↑**	**Acc.↑**	**H.Acc.↑**
Raw Model	0.342	0.450	0.516	0.605	0.503	0.677	0.623	0.561
Answer-only	0.562	0.662	0.479	0.560	0.497	0.659	0.523	0.427
Original Trace	0.426	0.518	0.521	0.608	0.512	0.667	0.627	0.535
STaR	0.510	0.613	0.613	0.689	0.583	0.745	0.587	0.541
NLL	0.546	0.642	0.600	0.693	0.597	0.767	0.617	0.554
RSR	0.568	0.668	0.624	0.720	0.600	0.765	0.577	0.490
LARK	0.539	0.626	0.630	0.712	0.607	0.774	0.603	0.522
MBT-R	0.579	0.672	0.633	0.719	0.627	0.768	0.620	0.567
REDIST (Ours)	**0.584**	**0.681**	**0.686**	**0.780**	**0.633**	**0.795**	**0.640**	**0.573**

4 EXPERIMENTS AND RESULTS

We evaluate whether REDIST turns student-generated attempts into supervision that improves both source-task performance and cross-benchmark transfer. Our experiments address three questions: (i) how REDIST compares with answer-only, original-trace, and existing trace-selection and rewriting methods, (ii) what trace construction and answer-loss allocation contribute to the resulting performance, and (iii) whether the gains extend across target students and how the choice of rewrite teacher affects downstream performance.

4.1 EXPERIMENTAL SETUP

Tasks and datasets. We use MuSiQue (Trivedi et al., 2022) as the primary source benchmark for training and source-task evaluation. We evaluate cross-benchmark performance on 2Wiki (Ho et al., 2020) and HotpotQA (Yang et al., 2018), which provide different multi-hop questions and evidence distributions, and on QuALITY (Pang et al., 2022), a long-context multiple-choice reading-comprehension benchmark. We report exact match (EM) and token-level F1 for MuSiQue, 2Wiki, and HotpotQA, and accuracy and hard-subset accuracy for QuALITY.

Baselines. *Raw Model* evaluates the student before task-specific training, *Answer-only* supervises the verified final answer, and *Original Trace* distills unmodified answer-correct student traces. We also compare with STaR-style rationalization (Zelikman et al., 2022), student-side trace selection using NLL, RSR (Yang et al., 2026), and LARK (Yu et al., 2026), and metacognitive rewriting with MBT-R (Kim et al., 2026). These comparisons assess the complete recipe against answer supervision, unmodified traces, trace selection, and trace rewriting. The implementation details are reported in Appendix B.3.

Models and training. Our primary experiments use Qwen3-8B (Yang et al., 2025a) as the target student and MuSiQue as the supervision source. Cross-student experiments additionally evaluate Qwen3-4B, Gemma4-E2B, and Gemma4-E4B (Gemma Team, 2026), with additional results provided in Table 3. REDIST performs trace construction and target selection offline, and inference uses only the distilled student. Full construction and optimization details are provided in Appendix B.2.

Evaluation protocol. The primary and cross-student comparisons use MuSiQue supervision, and a separate Qwen3-8B comparison uses 2Wiki as the training source (Table 7 in Appendix C). All REDIST results use $\alpha = 0.15$ and $K_d = K_r = 4$ unless noted otherwise. Table 9 in Appendix C reports results for other values of α. The primary comparison uses 1,000-example evaluation slices. Table 8 in Appendix C reports 95% bootstrap intervals over the evaluation slice for all Qwen3-8B point estimates .

4.2 SOURCE-TASK PERFORMANCE AND TRANSFER

Table 1 compares Qwen3-8B models trained with MuSiQue supervision and evaluated on MuSiQue, 2Wiki, HotpotQA, and QuALITY. All gradient updates use MuSiQue examples. REDIST achieves

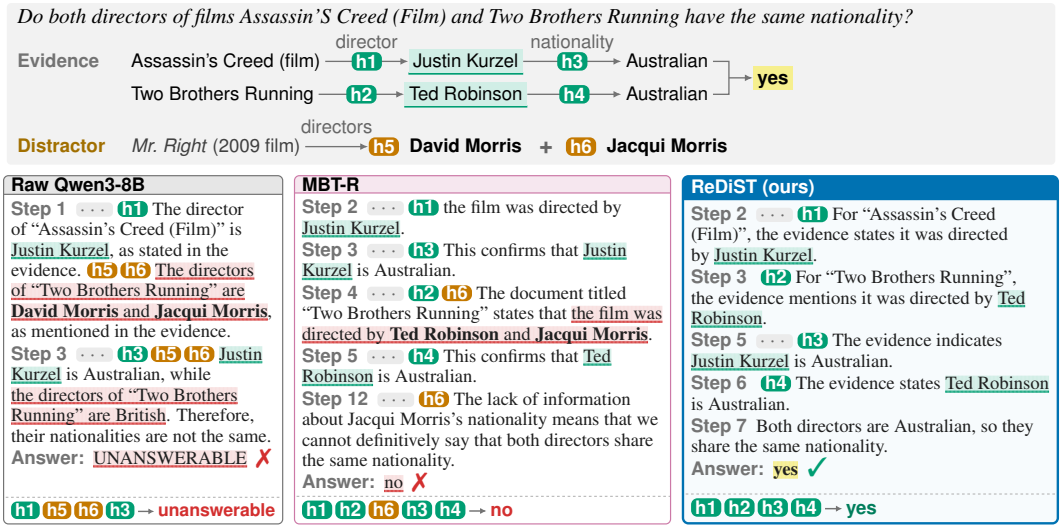


Figure 3: **Evidence use in a selected 2Wiki example.** Test-time reasoning excerpts from raw Qwen3-8B and Qwen3-8B students trained with MBT-R or REDIST. MBT-R identifies the relevant directors but introduces an unsupported co-director. In contrast, the REDIST output uses the relevant film–director–nationality links and answers correctly. Quoted model text is reproduced verbatim, with ellipses marking omissions, while highlights, correctness markers, and notes are author annotations.

the highest reported scores across all eight metrics, combining strong source-task performance with improvements over the evaluated supervision baselines on all three transfer benchmarks.

Source-task gains do not guarantee cross-benchmark gains. The answer-only and original-trace recipes illustrate this distinction. Answer-only training raises MuSiQue EM from 0.342 to 0.562, while 2Wiki EM falls from 0.516 to 0.479. Original-trace supervision reaches 0.426 and 0.521, respectively, reversing the ranking of these two recipes across benchmarks. Answer-only supervision directly trains prediction of the verified answer from the question and evidence, whereas trace supervision additionally makes intermediate statements and relations explicit learning targets. Such targets may convey evidence-linking structure that is useful on new questions, but their presence alone does not ensure effective distillation, as the weaker source-task performance of original traces illustrates. REDIST improves both source and transfer performance over these references.

QuALITY provides a complementary test of transfer to long-document, multiple-choice comprehension. Whereas most trace-based recipes improve 2Wiki and HotpotQA EM over the raw model, most do not improve its QuALITY performance. The raw model already reaches 0.623 accuracy and 0.561 hard-subset accuracy, while answer-only training reduces these scores to 0.523 and 0.427, respectively. REDIST reaches 0.640 and 0.573, modestly improving both raw-model references. Its gains on multi-hop benchmarks are therefore accompanied by preservation and modest improvement of performance on this more distinct transfer task. This pattern highlights the importance of evaluating both gains on related benchmarks and possible degradation on other question-answering tasks.

Similar source accuracy, stronger 2Wiki transfer. Among the trace-construction baselines, MBT-R is the closest comparison because it also rewrites student attempts and uses answer guidance to correct unsuccessful reasoning. MBT-R and REDIST achieve similar MuSiQue EM (0.579 vs. 0.584), but differ substantially on 2Wiki (0.633 vs. 0.686), a 5.3-point gap. The difference is smaller on HotpotQA (0.627 vs. 0.633), showing that the transfer advantage depends on the benchmark. Table 7 in Appendix C extends this comparison to Qwen3-8B trained with 2Wiki supervision. REDIST reaches 0.727 source EM and 0.536 MuSiQue transfer EM, compared with 0.707 and 0.512 for MBT-R. It also achieves 0.660 QuALITY accuracy versus 0.623 for MBT-R. These results extend the source-task and cross-benchmark gains beyond MuSiQue supervision, including the reverse 2Wiki-to-MuSiQue transfer direction.

Relevant facts can coexist with unsupported relations. In the selected 2Wiki example in Figure 3, MBT-R identifies the relevant directors and their nationalities but adds a co-director from a distractor film. Its trace then treats that person's unestablished nationality as relevant and answers incorrectly.

Table 2: **REDIST component and supervision-allocation ablations.** Variants modify trace construction or answer-loss allocation relative to the full recipe. All entries are EM.

		Source	Transfer	
Variant	Controlled change	MuSiQue	2Wiki	HotpotQA
Full REDIST	Reference recipe	0.584	**0.686**	**0.633**
ORIGINAL TARGET	No revision or repair	0.426	0.521	0.512
SELECTED STUDENT TARGET	No revision or repair	0.552	0.577	0.550
FRESH REWRITE	Rewrite without the original draft	0.554	0.483	0.570
REPAIR DISABLED	Original targets for repair cases	0.577	0.626	0.576
NO ANSWER AUXILIARY	$\alpha = 0$ in Eq. (4)	**0.592**	0.652	0.630
ANSWER-LOSS ONLY	$\alpha = 1$ in Eq. (4)	0.573	0.417	0.480
TRACE-ONLY CONTROL	Mask answer positions in loss	0.404	0.478	0.443

In contrast, the REDIST-trained student restricts the comparison to the two queried directors and answers correctly. *This case illustrates why process assessment should check unsupported additions alongside the required evidence links.*

These quantitative results compare complete training recipes, whereas the following component comparisons further examine the contributions of trace construction and answer-loss allocation.

4.3 TRACE CONSTRUCTION AND SUPERVISION ALLOCATION

Table 2 examines two decisions in REDIST: which intermediate targets to distill and how to allocate supervision to those targets. We first compare alternative trace constructions, then examine answer-loss allocation with the selected targets held fixed. Appendix B.4 details the controls used for these comparisons.

Construction adds gains beyond selecting student drafts. Compared with ORIGINAL TARGET, SELECTED STUDENT TARGET raises MuSiQue EM from 0.426 to 0.552 (+12.6 points) and also improves both transfer benchmarks. Full REDIST improves further over SELECTED STUDENT TARGET: MuSiQue EM rises from 0.552 to 0.584 (+3.2 points), while 2Wiki EM rises from 0.577 to 0.686 (+10.9 points) and HotpotQA EM from 0.550 to 0.633 (+8.3 points). Thus, *choosing among existing attempts already improves supervision, while the full construction pipeline yields additional gains that are larger on the transfer benchmarks.* This supports using the student's attempts as starting points for constructing teaching targets beyond those it initially generates.

Draft conditioning and repair matter most for transfer. Compared with full REDIST, FRESH REWRITE omits the student draft from rewrite input while retaining the same evidence and branch-specific answer access. MuSiQue EM falls from 0.584 to 0.554, 2Wiki EM from 0.686 to 0.483, and HotpotQA EM from 0.633 to 0.570. The largest decline is 20.3 points on 2Wiki. Fresh rewriting also falls below SELECTED STUDENT TARGET on 2Wiki (0.483 vs. 0.577), despite similar MuSiQue EM (0.554 vs. 0.552). *Conditioning the rewrite on the student's draft therefore yields better supervision than generating a new derivation.* The advantage of draft-conditioned construction supports using the student's existing reasoning path to inform what the teacher revises, even when substantial changes are allowed.

Comparing full REDIST with REPAIR DISABLED reveals a similar transfer-oriented pattern. Disabling repair reduces MuSiQue EM from 0.584 to 0.577, while 2Wiki EM falls from 0.686 to 0.626 and HotpotQA EM from 0.633 to 0.576. The repair-disabled control retains the same questions and uses original-target fallbacks for repair cases. With question coverage preserved, this comparison supports improving the supervision attached to initially unsuccessful attempts. These effects are less visible in source-task EM, making transfer evaluation particularly informative for draft conditioning and repair.

Answer weighting matters even when the traces are fixed. The three allocation variants at the bottom of Table 2 train on exactly the training mixture of full REDIST; they differ only in which supervised positions carry loss and how they are weighted. NO ANSWER AUXILIARY sets $\alpha = 0$ in Equation (4), so answer positions are weighted like any other position, and reaches 0.592 MuSiQue EM, 0.652 2Wiki EM, and 0.630 HotpotQA EM. The full recipe with $\alpha = 0.15$ is slightly lower on

Table 3: **Source-task and transfer performance across target students.** All trained variants receive MuSiQue supervision. Columns report MuSiQue source-task (S) and 2Wiki transfer (T) EM for each student.

	Qwen3-4B		Qwen3-8B		Gemma4-E2B		Gemma4-E4B	
Supervision	S	T	S	T	S	T	S	T
Raw model	0.293	0.225	0.342	0.516	0.225	0.518	0.425	0.643
Answer-only	0.558	0.440	0.562	0.479	0.514	0.368	0.578	0.449
Original trace	0.405	0.445	0.426	0.521	0.311	0.546	0.456	0.644
STaR	0.519	0.480	0.510	0.613	0.444	0.588	0.553	0.682
NLL	0.519	**0.635**	0.546	0.600	0.545	0.610	0.571	0.682
RSR	0.533	0.587	0.568	0.624	0.542	0.607	0.551	0.680
LARK	0.507	0.626	0.539	0.630	0.539	0.600	0.559	0.665
MBT-R	0.556	0.625	0.579	0.633	0.542	0.580	0.606	0.645
REDIST (Ours)	**0.580**	0.630	**0.584**	**0.686**	**0.559**	**0.622**	**0.608**	**0.723**

MuSiQue (0.584), essentially equal on HotpotQA (0.633), and 3.4 points higher on 2Wiki (0.686); Table 9 in Appendix C shows that larger values of α lower transfer EM. ANSWER-LOSS ONLY takes $\alpha = 1$, so only answer positions carry loss. It retains most of the source-task performance (0.573), but transfer falls to 0.417 on 2Wiki and 0.480 on HotpotQA, below the raw model. TRACE-ONLY CONTROL masks the answer positions out of the loss, and EM falls to 0.404, 0.478, and 0.443. Because long traces already leave the answer with a small share of a uniformly averaged objective, these results indicate that both parts of the completion need loss, with the answer share small but explicit: dropping answer supervision is harmful, supervising the answer alone forfeits the transfer gains that trace supervision provides, uniform averaging matches or slightly exceeds the full recipe on the source task, and moderate reweighting mainly benefits transfer to 2Wiki. Trace construction determines the intermediate supervision available to the student, while answer weighting controls the emphasis placed on predicting the verified final answer. Together, these comparisons support evaluating both the construction of intermediate targets and the training objective used to learn from them, with the largest gains over answer supervision appearing on the transfer benchmarks.

4.4 GENERALITY ACROSS STUDENT AND TEACHER MODELS

We evaluate REDIST across four target students and conduct a separate sweep of rewrite teachers. These experiments examine whether its benefits extend beyond the primary student and how teacher choice affects the resulting supervision. All settings use MuSiQue as the source of training examples.

Gains extend across target students. Table 3 compares four students from the Qwen3 and Gemma4 families. REDIST exceeds answer-only and original-trace supervision on both benchmarks for every student. It achieves the highest reported MuSiQue EM in all four comparisons and the highest 2Wiki EM in three. Beyond the primary Qwen3-8B comparison, Gemma4-E4B reaches 0.723 on 2Wiki, compared with 0.682 for the strongest alternatives, STaR and NLL, while Gemma4-E2B reaches 0.622, compared with 0.610 for NLL. Qwen3-4B is the exception: REDIST reaches 0.630, slightly below NLL at 0.635. *These results support the applicability of the construction and distillation pipeline across the evaluated model families and scales, while showing that its advantage over competing strategies depends on the target student.*

The source-transfer distinction also persists across students. Answer-only supervision outperforms original traces on MuSiQue but underperforms them on 2Wiki for all four students. For Gemma4-E4B, for example, answer-only supervision achieves higher source EM than original traces (0.578 vs. 0.456), but lower transfer EM (0.449 vs. 0.644). Thus, the different supervision profiles observed in Table 1 extend beyond Qwen3-8B, *reinforcing the value of evaluating supervision on both source and transfer distributions when the learner changes.*

Similar repair success can accompany different transfer outcomes. A separate rewrite-teacher sweep with Qwen3-8B student initialization reinforces the distinction between construction-time signals and downstream performance (Table 6 in Appendix C): Gemma4-26B-A4B has higher reported repair success than Qwen3-14B (62.15% versus 61.78%) but yields lower 2Wiki EM (0.624 vs. 0.686). This supports evaluating rewrite teachers through the distilled student's performance on the intended evaluation distributions.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, software implementation and debugging, experimental scripts, figure editing, literature identification, reference formatting, and feedback on the presentation and evaluation of experimental comparisons. Language models also serve as students, rewrite models, and readers within REDIST, where they generate and score constructed training traces used as synthetic training data as part of the studied method; this use is described in Sections 3 and 4. All AI-assisted code, calculations, figures, and manuscript edits were reviewed, tested, and verified by the authors before incorporation. The authors take responsibility for the final content of this work, including all reported results and AI-assisted artifacts.

ETHICS STATEMENT

This work studies reasoning-trace distillation on publicly available question-answering benchmarks (MuSiQue, 2Wiki, HotpotQA, and QuALITY), with models from the Qwen3 and Gemma4 families serving as students, rewrite models, and student-matched readers. It does not involve human subjects or the collection of personal data, and we use the datasets and models under their respective licenses. Distilled students can still produce incorrect answers or plausible-looking reasoning that is unsupported by the evidence, so their outputs should be independently verified before use in high-stakes settings.

REPRODUCIBILITY STATEMENT

Section 3 specifies the construction, selection, and training procedure, and Algorithm 1 in Appendix B.2 summarizes it. Appendix B gives the implementation details, including draft generation, validity rules, target selection, training exposure and loss, and optimization hyperparameters, as well as the definitions of the component ablations. Section 4.1 describes the benchmarks, evaluation slices, and metrics, and Appendix C reports the rewrite-teacher sweep, the comparison under 2Wiki supervision, and the answer-loss weight sweep. All benchmarks and model families used in this work are publicly available.

REFERENCES

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Edoardo Cetin, Tianyu Zhao, and Yujin Tang. Reinforcement Learning Teachers of Test Time Scaling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 119028–119062, 2025.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. Re-ReST: Reflection-Reinforced Self-Training for Language Agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15394–15411, 2024.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. TRACE the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8472–8494, 2024.

Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

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, 2023.

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.

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. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Jinghan Jia, Joe Benton, and Eric Easley. Faithfulness as Information Flow: Evaluating and Training Faithful Chain-of-Thought Reasoning. *arXiv preprint arXiv:2605.24286*, 2026.

Berkcan Kapusuzoglu, Supriyo Chakraborty, Zain Sarwar, Chia-Hsuan Lee, and Sambit Sahu. Critique-Guided Distillation for Robust Reasoning via Refinement. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Ik-hwan Kim, Hyeongrok Han, Mingi Jung, Sangwon Yu, Jinseok Hong, Sang Hun Kim, Yoonyoung Choi, and Sungroh Yoon. Metacognitive Behavioral Tuning of Large Language Models for Multi-Hop Question Answering. *arXiv preprint arXiv:2602.22508*, 2026.

Minsang Kim and Seung Jun Baek. Explain in Your Own Words: Improving Reasoning via Token-Selective Dual Knowledge Distillation. In *International Conference on Learning Representations*, 2026.

Siqi Kou, Qingyuan Tian, Hanwen Xu, Zihao Zeng, and Zhijie Deng. Which Data Attributes Stimulate Math and Code Reasoning? An Investigation via Influence Functions. In *Advances in Neural Information Processing Systems*, volume 38, pp. 13789–13811, 2025.

Jungseob Lee, Seungyoon Lee, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, and Heuiseok Lim. Answer-Conditioned Chains of Thought Degrade Verifiable-Reasoning Distillation in Large Language Models. *arXiv preprint arXiv:2607.14552*, 2026.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and Xian Li. NaturalThoughts: Selecting and Distilling Reasoning Traces for General Reasoning Tasks. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025.

Fabio Valerio Massoli, Andrey Kuzmin, and Arash Behboodi. Reasoning as Compression: Unifying Budget Forcing via the Conditional Information Bottleneck. In *1st Workshop on Scaling Post-Training for LLMs at ICLR 2026*, 2026.

Ethan Mendes, Jungsoo Park, and Alan Ritter. Making Expert Reasoning Learnable with Self-Distillation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question Answering with Long Input Texts, Yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, 2022.

Sohan Patnaik, Milan Aggarwal, Sumit Bhatia, and Balaji Krishnamurthy. It Helps to Take a Second Opinion: Teaching Smaller LLMs To Deliberate Mutually via Selective Rationale Optimisation. In *International Conference on Learning Representations*, 2025.

Haolong Qian, Xianliang Yang, Yinuo Ma, Lirong Che, Feng Lu, Ye Guo, Lei Song, Jiang Bian, and Chun Yuan. The Quality–Utility Paradox: Why High-Reward Data Impairs Small Model Mathematical Reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023.

Xiangyu Wen, Junhua Huang, Zeju Li, Min Li, Jianyuan Zhong, Zhijian Xu, Mingxuan Yuan, Yongxiang Huang, and Qiang Xu. Reasoning Scaffolding: Distilling the Flow of Thought from LLMs. In *International Conference on Learning Representations*, 2026.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A Theory of Usable Information under Computational Constraints. In *International Conference on Learning Representations*, 2020. Oral presentation.

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

Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E. Gonzalez, Bin Cui, and Shuicheng Yan. SuperCorrect: Advancing Small LLM Reasoning with Thought Template Distillation and Self-Correction. In *International Conference on Learning Representations*, 2025b.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, Yunhua Zhou, Yicheng Zou, Qipeng Guo, Tao Gui, Qi Zhang, and Xuanjing Huang. Which Reasoning Trajectories Teach Students to Reason Better? A Simple Metric of Informative Alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42123–42150, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *International Conference on Learning Representations*, 2024.

Tianrun Yu, Kaixiang Zhao, Chih-Chun Chen, Amanda Hughes, Taylor W. Killian, Fenglong Ma, Weitong Zhang, and Porter Jenkins. LARK: Learnability-Grounded Trajectory Selection for Efficient Reasoning Distillation. *arXiv preprint arXiv:2605.30651*, 2026.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022.

Haoran Zhang, Yafu Li, Zhi Wang, Zhilin Wang, Shunkai Zhang, Xiaoye Qu, and Yu Cheng. Characterizing, Evaluating, and Optimizing Complex Reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a.

Ruichen Zhang, Rana Muhammad Shahroz Khan, Zhen Tan, Dawei Li, Song Wang, and Tianlong Chen. The Quest for Efficient Reasoning: A Data-Centric Benchmark to CoT Distillation. In *International Conference on Learning Representations*, 2026b.

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 *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a.

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.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. ReasonFlux-PRM: Trajectory-Aware PRMs for Long Chain-of-Thought Reasoning in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 30320–30358, 2025.

A Related Work

Trace construction and self-training. One line of work improves reasoning supervision by generating, correcting, or restructuring the traces used for training. STaR bootstraps reasoning from model-generated rationales and uses answer-conditioned rationalization when initial attempts fail (Zelikman et al., 2022). Re-ReST repairs low-quality trajectories using external feedback (Dou et al., 2024), and SuperCorrect and CGD incorporate correction into distillation (Yang et al., 2025b; Kapusuzoglu et al., 2026). Metacognitive Behavioral Tuning rewrites or synthesizes traces with an explicit multi-stage metacognitive structure (Kim et al., 2026). Its MBT-R variant also rewrites student attempts and uses answer guidance to correct unsuccessful reasoning, making it our closest baseline. Controlled work shows that answer-conditioned rationalization can degrade verifiable-reasoning distillation even when the final answer is correct (Lee et al., 2026). REDIST constructs teaching targets by editing the target student's own attempts and gives the verified answer only to the repair of unsuccessful drafts. Our fresh-rewrite control indicates that this draft conditioning matters most for transfer (Section 4.3).

Trajectory selection and learner compatibility. A second line of work asks which traces a particular student can learn from, and selects, scores, or adapts trajectories accordingly. RSR measures student-specific informative alignment through rank and surprisal (Yang et al., 2026), and LARK selects trajectories using a learnability proxy and distributional coverage (Yu et al., 2026). The Quality–Utility Paradox shows that higher reward-model quality can be offset by learner–data mismatch and addresses this mismatch through style-preserving refinement (Qian et al., 2026). DAIL and TSD-KD address learner–teacher mismatch (Mendes et al., 2026; Kim & Baek, 2026). RLT and COALITION use student answer prediction to optimize explanations or select rationales (Cetin et al., 2025; Patnaik et al., 2025), while TRM and ReasonFlux-PRM score reasoning quality (Zhang et al., 2026a; Zou et al., 2025). REDIST first creates new candidates by editing the student's own drafts and then chooses among the original and edited versions with a frozen student-matched reader, using this student signal offline without training the rewrite model.

Distillation objectives and training data. Distillation outcomes also depend on how traces enter training and on the composition of the training data. Distilling step-by-step trains rationale generation and label prediction as separate tasks (Hsieh et al., 2023). OPSD distills a privileged self-teacher into on-policy student rollouts, ROSD adds reflection-guided, error-localized correction, and SDPO turns textual feedback into dense self-distillation signals during reinforcement learning (Zhao et al., 2026a;b; Hübotter et al., 2026). Reasoning Scaffolding distills reasoning structure (Wen et al., 2026), and influence-based attribution relates training data to reasoning gains (Kou et al., 2025). NaturalThoughts studies the scale, difficulty, and diversity of distilled reasoning data and its mixture with answer-only data (Li et al., 2025), and DC-CoT benchmarks data selection and mixing for chain-of-thought distillation and their effect on generalization (Zhang et al., 2026b). Unlike on-policy self-distillation, REDIST constructs targets offline and trains with supervised fine-tuning on a single completion containing the trace and the verified answer.

Evidence grounding and process measurement. Our setting also connects to work on reasoning over supplied evidence and on what visible reasoning traces reveal. In retrieval-augmented and evidence-rich question answering, prior work trains models to be robust to irrelevant context (Yoran et al., 2024) and constructs knowledge-grounded reasoning chains from retrieved documents (Fang et al., 2024). Chain-of-thought explanations need not faithfully reflect the computation behind a prediction (Turpin et al., 2023; Arcuschin et al., 2025), and information-bottleneck and information-flow views analyze reasoning compression and faithfulness (Massoli et al., 2026; Jia et al., 2026). Predictive $\mathcal{V}$-information formalizes the information that a restricted computational family can extract (Xu et al., 2020). REDIST uses the supplied evidence to revise and repair the student's own reasoning before distillation and treats visible traces only as teaching targets.

B Experimental and Implementation Details

This appendix specifies the construction, training, and evaluation details supporting the main comparisons.

B.1 TASKS, STUDENTS, AND COMPARISON ENDPOINTS

The primary comparison trains Qwen3-8B with MuSiQue supervision and evaluates source-task performance and transfer to 2Wiki (Ho et al., 2020), HotpotQA (Yang et al., 2018), and QuALITY (Pang et al., 2022). The model sweep varies the target student under MuSiQue supervision. A separate second-source study trains Qwen3-8B on 2Wiki and reports source-task and transfer outcomes. The rewrite-teacher sweep varies the rewrite model with a fixed Qwen3-8B student initialization. These studies separate the effects of changing the learner, the training source, and the rewrite teacher.

QA results use answer EM and token F1, and QuALITY results use accuracy and hard-subset accuracy. Each metric is interpreted within its own task. The raw model anchors performance before task-specific training, and each student/task setting has its own answer-only reference for Equation (1).

B.2 REDIST IMPLEMENTATION

Draft generation and selection. For each training example (Q, E, Y), REDIST first generates K_d candidate drafts from the target student. The implementation uses one deterministic draft followed by sampled drafts. Unless otherwise specified, the frozen reader is a copy of the target student. A draft is reader-eligible only when it contains a complete nonempty reasoning trace and a well-formed final answer. For an eligible trace R, the reader is given the student prompt followed by R. Its answer prediction $\hat{Y}^{\text{read}}_{s_0}(Q, E, R)$ is the greedy continuation from this position up to answer closing tag, and its verified-answer score $q_{s_0}(Y \mid Q, E, R)$ is the product of the reader's next-token probabilities over the tokens of Y. REDIST first restricts attention to drafts whose generated final answer matches Y when such drafts exist, and otherwise considers all reader-eligible drafts. In either case, it selects the candidate with the largest reader score. Ties are resolved deterministically by candidate order. If no draft is reader-eligible, the implementation falls back to the first generated draft for deterministic case classification.

Revision and repair. The selected draft determines which rewrite prompt is used. For an answer-correct draft, the rewrite model receives only the question, evidence, and draft reasoning. For an unsuccessful draft, the verified answer is additionally supplied as a checking target for answer-assisted repair. In both branches, the rewrite model is instructed to produce a complete evidence-grounded explanation, retain useful supported reasoning, remove unsupported or irrelevant content, and explicitly connect the relevant intermediate entities and relations. The rewrite is allowed to reorder, replace, add, or remove steps and to change the wording, so REDIST conditions on an existing draft without imposing a minimal-edit constraint. Each branch generates K_r rewrite candidates. Empty or length-capped generations, traces containing reasoning/answer tags, and direct answer-declaration patterns are removed by deterministic validity filters.

Student-conditioned target selection. The original draft and all valid rewrite candidates are rescored by the frozen reader. A candidate passes the functional answer-recovery check when the reader's greedy answer $\hat{Y}^{\text{read}}_{s_0}(Q, E, R)$ matches Y under the normalized exact-match rule used for evaluation. If at least one candidate passes, REDIST selects the passing candidate with the largest $q_{s_0}(Y \mid Q, E, R)$. If none passes and the original draft was answer-correct, the original draft is retained. If none passes for an unsuccessful draft, the target becomes answer-only. A selected generated candidate from the unsuccessful branch is labeled as a repair, while a selected original or correct-branch candidate is labeled as retained or revised. The answer-recovery check provides a functional selection criterion.

Validity and format-only cases. Before rewrite and training, deterministic parsing distinguishes ordinary unsuccessful drafts from answer-like or format-only cases. The current implementation excludes these format-only cases from training. Controlled ablations define their shared cohort after this deterministic preprocessing step so that branch-specific interventions do not change the training question set.

Optimization. The REDIST SFT implementation uses rank-32 LoRA (Hu et al., 2022) with LoRA scaling 64 and dropout 0.05 on the attention projection modules and the MLP gate/up/down

Algorithm 1 REDIST trace construction and distillation

Require: Training data $\mathcal{D}$; initial student s_0; teacher g; frozen student-matched reader; draft/edit counts K_d, K_r; repair multiplicity m; answer-loss weight α
Ensure: Distilled student $s^\star$

$\widetilde{\mathcal{D}} \leftarrow \emptyset$
for all $(Q, E, Y) \in \mathcal{D}$ **do**
  $\mathcal{A}_{\text{draft}} \leftarrow \textsc{GenerateAttempts}(s_0, Q, E, K_d)$
  $a_0 \leftarrow \textsc{SelectDraft}(\mathcal{A}_{\text{draft}}, Q, E, Y, q_{s_0})$
  if $a_0 = \perp$ **then**
    continue ▷ Documented parsing exclusions
  end if
  $(R^{(0)}, \widehat{Y}^{(0)}) \leftarrow a_0$
  $b \leftarrow [\widehat{Y}^{(0)} \text{ matches } Y]$
  if b **then**
    $\mathcal{R}_{\text{new}} \leftarrow \textsc{Revise}(g, Q, E, R^{(0)}, K_r)$
  else
    $\mathcal{R}_{\text{new}} \leftarrow \textsc{Repair}(g, Q, E, R^{(0)}, Y, K_r)$
  end if
  $\mathcal{C} \leftarrow \textsc{BuildCandidates}(R^{(0)}, \mathcal{R}_{\text{new}}, b)$
  $(R^\star, \tau^\star) \leftarrow \textsc{SelectRecoverable}(\mathcal{C}, a_0, Q, E, Y)$
  $w \leftarrow m$ if $\tau^\star = \textsc{Repair}$, else $w \leftarrow 1$
  Append w copies of $(Q, E, R^\star, Y)$ to $\widetilde{\mathcal{D}}$
end for
$s^\star \leftarrow \textsc{Distill}(s_0, \widetilde{\mathcal{D}}, \alpha)$ using Equation (4)
return $s^\star$

projections. Training uses AdamW with learning rate 10^{-4}, minibatch size 2, gradient accumulation over 8 minibatches, cosine learning-rate decay with a short warmup, gradient clipping at norm 1, and gradient checkpointing. Base-model parameters are loaded in bfloat16 and frozen outside the LoRA adapters. All runs use 500 optimizer steps with the same batch size. Each mini-batch is drawn from the method's own training mixture. Except for MBT-R, we use a generation limit of 1,024 tokens. We use seed 42 to randomly sample 4,000 training examples from the full training set. Unless otherwise specified, we set $K_d = K_r = 4$, $m = 4$, and $\alpha = 0.15$. Unless noted otherwise, the rewrite model is Qwen3-14B for Qwen students and Gemma4-26B-A4B for Gemma students, and the frozen reader is a copy of the target student.

Algorithm summary. In Algorithm 1, SELECTDRAFT follows the draft-selection rules in Section 3.1 and returns $\perp$ only for the documented format-only exclusions. BUILDCANDIDATES applies the validity rules and preserves candidate origins. SELECTRECOVERABLE uses the frozen reader to apply Equation (3) and returns the selected trace $R^\star$ with its origin $\tau^\star$, including the original-draft and answer-only fallbacks.

B.3 BASELINE IMPLEMENTATIONS

All baselines use the same target student, the same 4,000 training questions, the same LoRA and optimizer configuration, the same number of optimizer steps, and the same evaluation setting as REDIST. They differ mainly in the supervision target attached to each training question, and therefore in which questions are retained. We group the baselines by where the supervision target comes from (Table 4): a *reference* without task training, *answer supervision*, traces *self-generated* by the student, traces *selected* from a teacher candidate pool by a student-side score, and traces *rewritten* by a teacher.

Shared protocol. The training source is a fixed set of 4,000 questions drawn uniformly without replacement (seed 42) from the training split. Every question keeps its full evidence set, with supporting and distractor paragraphs and no truncation. All methods start from these 4,000 questions. Original Trace and STaR retain questions according to their own acceptance rules. The trace-selection arms retain the questions that have at least one usable teacher candidate, and Answer-only trains on all

Table 4: **Baseline taxonomy.** "Student" denotes the untrained target student. The teacher is Qwen3-14B for Qwen students and Gemma4-26B-A4B for Gemma students.

Category	Baseline	Trace source	Acceptance / selection
Reference	Raw Model	— (no training)	—
Answer supervision	Answer-only	— (gold answer only)	aligned question set
Self-generated	Original Trace	student, greedy	EM match, non-empty trace
	STaR	student, greedy (+ hint)	EM match, non-empty trace, closed tag (hint removed for training)
Trace selection	NLL	teacher, multi-prompt pool	arg min mean student NLL
	RSR	teacher, multi-prompt pool	arg min rank–surprisal ratio
	LARK	teacher, multi-prompt pool	arg max LARK score
Trace rewriting	MBT-R	teacher rewrite of student draft	none (one rewrite per question)

4,000 questions; MBT-R rewrites every parseable draft under MuSiQue supervision and is restricted to the selection cohort under 2Wiki supervision. Every table entry evaluates the final checkpoint of its run on the first 1,000 examples of the validation split, and QuALITY entries use a fixed 300-question set.

Output format. Trace-supervised targets place the reasoning in a `<think>...</think>` block, followed by `<answer>` gold answer `</answer>` and the end-of-sequence token. Qwen students are prompted with a raw-text instruction that ends inside an open `<think>` block, so their native chat-template thinking mode is not used. Gemma students use their chat template with thinking disabled. The answer span of every training target carries the dataset's gold string, regardless of the answer written in the trace. Every Qwen prompt, including the raw, answer-only, and trace prompts, also allows the answer *UNANSWERABLE* when the evidence is insufficient, and such answers are scored as incorrect. The Gemma prompts have no abstention option, so Raw and Answer-only values should be compared within a model family.

Evaluation. Decoding is greedy. Exact match (EM) and token-level F1 are computed after standard answer normalization, taking the maximum over the gold answer and its annotated aliases, over all 1,000 examples of each benchmark. Tagged outputs are scored on the content of the first closed `<answer>...</answer>` span. Outputs without a closed answer span are scored on their first non-empty line. For Qwen students, this line is the opening `<think>` tag and scores zero. For Gemma students, it is the first generated line, which occasionally earns partial F1. Answer-only outputs are scored on their first non-empty line. The generation budget is 1,024 new tokens by default.

Methods with self-generated traces. Following STaR (Zelikman et al., 2022), questions the student already solves keep their native trace. For every rejected question, the same untrained student generates a rationalization with the gold answer appended to the question as a hint ("*(The correct answer is: ...)*"), using greedy decoding with a 768-token budget. A rationalization is accepted only if its own final answer matches the gold answer under normalized exact match, its reasoning block is non-empty, and its answer tag is closed. Questions that fail both passes are discarded. The hint is removed from the training prompt, and accepted rationalizations are otherwise used unchanged, so some of them refer to the provided answer.

Methods with teacher trace selection. NLL, RSR, and LARK select from one shared pool generated by a teacher from the student's own model family: Qwen3-14B for Qwen students and Gemma4-26B-A4B for Gemma students. The teacher answers every training question under several gold-blind prompt variants that differ only in the requested reasoning style. We use six variants (standard, decomposition, evidence chain, entity tracking, verification, minimal chain), each replacing the reasoning line of the standard prompt. Within each question, exact duplicate traces are collapsed to a single candidate. Each method follows the selection rules given in Table 4.

Trace rewriting methods. We follow the two-turn rewriting protocol of MBT-R (Kim et al., 2026) with its official prompt templates, using Qwen3-14B for Qwen students and Gemma4-26B-A4B for Gemma students as the rewriting teacher, so that all teacher-generated schemes share the same teacher.

Retained training data. Table 5 lists how many of the 4,000 sampled questions each method retains and what its training mixture contains. Answer-only keeps every question; Original trace keeps only answer-correct greedy drafts; STaR recovers rejected questions through hinted rationalization; the trace-selection arms keep questions with at least one usable teacher candidate; MBT-R rewrites every draft, except questions on which the teacher's rewrite falls into a repetition loop and exhausts the generation budget; and REDIST drops only the questions whose selected draft violates the requested output format. Under MuSiQue supervision, 1,004 of 1,625 unsuccessful drafts are repaired and 621 fall back to answer-only targets; under 2Wiki supervision, 1,089 of 1,371 are successfully repaired.

Table 5: **Training data retained by each supervision method (Qwen3-8B student).** Kept counts distinct training questions.

Method	Kept	Dropped	Training mixture
MuSiQue supervision (4,000 training questions)			
Answer-only	4,000	0	4,000 gold answers, no trace
Original trace	1,466	2,534	1,466 answer-correct greedy traces
STaR	3,671	329	1,466 native traces + 2,205 accepted rationalizations (of 2,529 attempted)
NLL / RSR / LARK	2,477	1,523	2,477 selected teacher traces, one per question
MBT-R	3,995	5	3,995 rewrites of the student draft
REDIST	3,691	309	1,813 revised + 253 retained correct drafts; 1004 accepted repairs; 621 answer-only fallbacks
2Wiki supervision (4,000 training questions)			
Answer-only	4,000	0	4,000 gold answers, no trace
Original trace	2,296	1,704	2,296 answer-correct greedy traces
STaR	3,895	105	2,296 native traces + 1,599 accepted rationalizations (of 1,704 attempted)
NLL / RSR / LARK	3,408	592	3,408 selected teacher traces, one per question
MBT-R	3,408	592	3,408 rewrites of the student draft (selection cohort)
REDIST	3,933	67	2,439 revised + 123 retained correct drafts; 1,089 accepted repairs; 282 answer-only fallbacks

B.4 ABLATION DETAILS

Table 2 compares full REDIST with different variants. ORIGINAL TARGET is the Original Trace baseline of Table 1, which trains on unmodified answer-correct student drafts. SELECTED STUDENT TARGET keeps the selected draft of every question as its target, without revision or repair. FRESH REWRITE gives the rewrite model the question, the evidence, and the same branch-specific answer access as REDIST, omits the student draft and its answer, and applies the same checks and fallbacks to the resulting candidates. REPAIR DISABLED uses the original draft as the target for every question whose selected draft was unsuccessful. These drafts can reach a wrong answer, while the supervised final answer remains the verified answer. In SELECTED STUDENT TARGET and REPAIR DISABLED, a question whose draft is empty receives an answer-only target. SELECTED STUDENT TARGET, FRESH REWRITE, and REPAIR DISABLED keep the repair exposure of full REDIST, so questions with accepted repairs in REDIST appear m times in training mixture even when their new targets are not repairs. NO ANSWER AUXILIARY, ANSWER-LOSS ONLY, and TRACE-ONLY CONTROL all train on exactly the training mixture of full REDIST, with the same selected targets, the same repair exposure m, the same tagged prompt and output format, and the same optimizer settings; only the loss allocation over supervised positions changes. NO ANSWER AUXILIARY sets $\alpha = 0$ in Equation (4), which reduces the objective to a uniform token-averaged cross-entropy over all completion positions. ANSWER-LOSS ONLY sets $\alpha = 1$, so only answer-content positions carry loss while the trace remains in the completion as conditioning context. TRACE-ONLY CONTROL computes the loss only over trace positions and excludes the answer-content positions, so answer tokens receive no direct supervision.

C Additional Results

Rewrite-teacher sweep. Table 6 varies the rewrite teacher while initializing each distilled student from Qwen3-8B. MuSiQue EM ranges from 0.578 to 0.593 across teachers, whereas 2Wiki EM ranges from 0.620 to 0.686. The highest reported repair success belongs to Gemma4-26B-A4B (62.15%), slightly above Qwen3-14B (61.78%), but their resulting students reach 0.624 and 0.686 2Wiki EM, respectively. The Qwen3-8B teacher gives the highest source-task EM at 0.593. Thus, repair success, source-task accuracy, and transfer performance favor different rewrite teachers. Teacher choice should be evaluated through the distilled student's performance on the intended evaluation distributions.

Token-level normalized edit distance (Token NED) is the minimum number of token insertions, deletions, and substitutions between the draft and edited trace, divided by the longer trace length and expressed as a percentage. It provides a complementary description of the two editing branches. Repair NED exceeds revision NED for every teacher, and for Qwen3-14B, the values are 29.90% and 1.15%, respectively. This pattern is consistent with different editing roles assigned to unsuccessful and answer-correct attempts. Yet similar repair-edit magnitudes can accompany different transfer outcomes: Qwen3-14B and Gemma4-26B-A4B have repair NED values of 29.90% and 30.04%, despite their different 2Wiki EM scores above. Edit magnitude characterizes how much the trace changes, while downstream evaluation assesses the usefulness of the resulting supervision.

Table 6: **Rewrite-teacher comparison with a fixed student initialization.** Each student is initialized from Qwen3-8B and distilled using MuSiQue supervision. We report repair success, token-level normalized edit distance (NED), and downstream EM/F1.

Rewrite teacher	Repair success	Token NED (%)		MuSiQue (Source)		2Wiki (Transfer)	
	(%)	Revise	Repair	EM	F1	EM	F1
Qwen3-4B	57.97	0.98	18.31	0.581	0.691	0.648	0.733
Qwen3-8B	61.60	2.78	26.24	**0.593**	0.693	0.670	0.757
Qwen3-14B	61.78	1.15	29.90	0.584	0.681	**0.686**	**0.780**
Qwen3-30B-A3B	53.29	5.92	28.22	0.587	**0.695**	0.646	0.729
Gemma4-E4B	56.43	13.10	44.86	0.581	0.686	0.670	0.763
Gemma4-12B	60.43	9.72	35.97	0.582	0.685	0.620	0.714
Gemma4-26B-A4B	**62.15**	8.27	30.04	0.578	0.689	0.624	0.722

Table 7: **Source-task and cross-benchmark performance after 2Wiki supervision.** All systems use Qwen3-8B. Trained variants use only 2Wiki examples for gradient updates.

	2Wiki (Source)		**MuSiQue** (Transfer)		**HotpotQA** (Transfer)		**QuALITY** (Transfer)	
Method	**EM↑**	**F1↑**	**EM↑**	**F1↑**	**EM↑**	**F1↑**	**Acc.↑**	**H.Acc.↑**
Raw Model	0.516	0.605	0.342	0.450	0.503	0.677	0.623	0.561
Answer-only	0.698	0.775	0.482	0.579	0.630	0.763	0.537	0.465
Original Trace	0.599	0.676	0.387	0.486	0.589	0.736	0.620	0.567
STaR	0.675	0.731	0.448	0.548	0.597	0.740	0.623	0.535
NLL	0.677	0.763	0.463	0.561	0.616	0.762	0.607	0.522
RSR	0.629	0.718	0.449	0.567	0.634	0.781	0.613	0.554
LARK	0.684	0.769	0.469	0.561	0.625	0.768	0.577	0.490
MBT-R	0.707	0.755	0.512	0.618	0.630	**0.782**	0.623	0.535
REDIST (Ours)	**0.727**	**0.791**	**0.536**	**0.633**	**0.638**	**0.782**	**0.660**	**0.580**

Bootstrap intervals. Table 8 reports 95% percentile bootstrap intervals for the REDIST Qwen3-8B point estimates, obtained by resampling the fixed 1,000-question evaluation slice with replacement 10,000 times. On a 1,000-question slice the intervals are roughly six points wide, which sets the resolution of the comparisons. On the source tasks, REDIST has the highest point estimate in both directions, but its interval contains the point estimates of the strongest baselines (MBT-R, RSR, and Answer-only under MuSiQue supervision; MBT-R under 2Wiki supervision), so the ordering among the top source-task recipes is not resolved at this sample size. The MuSiQue-to-2Wiki transfer margin is different in kind: no baseline point estimate falls inside the REDIST interval, and the 5.3-point

Table 8: **Qwen3-8B source-task and transfer EM under MuSiQue and 2Wiki supervision.** Each entry reports EM on a fixed 1,000-question validation slice with its 95% bootstrap interval.

	MuSiQue-trained		2Wiki-trained	
Supervision	MuSiQue (Source)	2Wiki (Transfer)	2Wiki (Source)	MuSiQue (Transfer)
Raw model	0.342 [0.313, 0.370]	0.516 [0.485, 0.547]	0.516 [0.485, 0.547]	0.342 [0.313, 0.370]
Answer-only	0.562 [0.531, 0.592]	0.479 [0.448, 0.510]	0.698 [0.669, 0.726]	0.482 [0.451, 0.513]
Original trace	0.426 [0.396, 0.457]	0.521 [0.490, 0.551]	0.599 [0.568, 0.629]	0.387 [0.357, 0.417]
STaR	0.510 [0.479, 0.542]	0.613 [0.583, 0.643]	0.675 [0.646, 0.704]	0.448 [0.417, 0.478]
NLL	0.546 [0.515, 0.577]	0.600 [0.570, 0.630]	0.677 [0.648, 0.705]	0.463 [0.433, 0.493]
RSR	0.568 [0.537, 0.599]	0.624 [0.594, 0.654]	0.629 [0.599, 0.659]	0.449 [0.418, 0.479]
LARK	0.539 [0.508, 0.569]	0.630 [0.600, 0.660]	0.684 [0.655, 0.712]	0.469 [0.438, 0.500]
MBT-R	0.579 [0.548, 0.609]	0.633 [0.602, 0.662]	0.707 [0.677, 0.736]	0.512 [0.479, 0.542]
REDIST (Ours)	**0.584** [0.553, 0.615]	**0.686** [0.658, 0.714]	**0.727** [0.700, 0.754]	**0.536** [0.505, 0.567]

gap over MBT-R exceeds the interval half-width. In the reverse direction, REDIST exceeds every baseline except MBT-R by more than the interval half-width, while the 2.4-point margin over MBT-R lies within it.

Answer-loss weight. Table 9 sweeps α with the selected targets and exposure of full REDIST held fixed; $\alpha = 0$ is the NO ANSWER AUXILIARY variant of Table 2. Source-task EM is highest at $\alpha = 0$ (0.592) and varies by less than one point up to $\alpha = 0.15$, whereas both transfer benchmarks peak at $\alpha = 0.15$ (0.686 on 2Wiki, 0.633 on HotpotQA). Beyond $\alpha = 0.15$, all three benchmarks decline, with 2Wiki EM falling monotonically to 0.579 at $\alpha = 0.90$. Considering the trade-off between different metrics, we use $\alpha = 0.15$ throughout.

Table 9: **Answer-loss weight sweep.**

α	**MuSiQue EM↑**	**2Wiki EM↑**	**HotpotQA EM↑**
0.00	**0.592**	0.652	0.630
0.15	0.584	**0.686**	**0.633**
0.30	0.572	0.653	0.602
0.60	0.575	0.616	0.587
0.75	0.558	0.610	0.590
0.90	0.535	0.579	0.567

D PROMPTS

In this section, we provide the prompts used in our experiments. The student prompt below is the raw-text form used for Qwen students, which ends inside an open `<think>` block. Gemma students receive the same instruction (except for the abstention option), question, and evidence as the user turn of their chat template with thinking disabled; the tagged output format is unchanged. The reader prompt is wrapped in the same way as the student prompt of the corresponding model family, and the teacher prompts are sent as a single user turn of the rewrite model's chat template.

REDIST Student Prompt (Draft Generation, Training, and Evaluation)

```
You are a precise multi-hop QA assistant.
Read the EVIDENCE and answer the QUESTION.
You MUST reason in multiple concise steps before final answer.
Output format must be EXACTLY:
<think>
1) step one
2) step two
...
</think>
```

```
<answer>short final answer phrase</answer>
No text outside these tags.
If the EVIDENCE is insufficient to determine the answer, give
↪  UNANSWERABLE as the final answer.

QUESTION:
<question>

EVIDENCE:
Title: <title_1>
<paragraph_1>

Title: <title_2>
<paragraph_2>
...

OUTPUT:
<think>
```

REDIST Teacher Rewrite Prompt (Correct Draft)

```
The student's draft reached the correct answer. Improve its clarity
↪  and evidential support without changing the supported
↪  conclusion.
Rewrite the reasoning into a complete evidence-grounded explanation.
↪  You may reorder, replace, add, or remove steps and change the
↪  wording; you do not need to minimize edits or preserve the
↪  original length.
Resolve the intermediate entities and explicitly connect the
↪  relevant facts from the evidence to the final answer-bearing
↪  fact. Keep useful reasoning, remove unsupported claims and
↪  irrelevant details. Use only the evidence, not outside
↪  knowledge. Prefer a concise complete chain, usually 3-7 numbered
↪  steps; do not pad easy questions.
State the final relationship as a natural factual statement. Do not
↪  write 'the answer is', repeat an answer hint, mention a
↪  supplied/gold answer, the draft, or the rewriting instructions.
↪  Output ONLY the reasoning, with no think/answer tags or separate
↪  answer declaration.

QUESTION:
<question>

EVIDENCE DOCUMENTS:
<evidence>

DRAFT REASONING:
<student_draft_reasoning>

REWRITTEN REASONING:
```

REDIST Teacher Repair Prompt (Incorrect Draft)

```
The student's draft reached a wrong answer. Correct the faulty
↪  reasoning and derive the supplied target from the evidence. The
↪  target is a checking hint, not evidence; do not invent facts to
↪  justify it.
Rewrite the reasoning into a complete evidence-grounded explanation.
↪  You may reorder, replace, add, or remove steps and change the
↪  wording; you do not need to minimize edits or preserve the
↪  original length.
```

```
Resolve the intermediate entities and explicitly connect the
↪  relevant facts from the evidence to the final answer-bearing
↪  fact. Keep useful reasoning, remove unsupported claims and
↪  irrelevant details. Use only the evidence, not outside
↪  knowledge. Prefer a concise complete chain, usually 3-7 numbered
↪  steps; do not pad easy questions.
State the final relationship as a natural factual statement. Do not
↪  write 'the answer is', repeat an answer hint, mention a
↪  supplied/gold answer, the draft, or the rewriting instructions.
↪  Output ONLY the reasoning, with no think/answer tags or separate
↪  answer declaration.

QUESTION:
<question>

EVIDENCE DOCUMENTS:
<evidence>

DRAFT REASONING:
<student_draft_reasoning>

CORRECT FINAL ANSWER:
<gold_answer>

REWRITTEN REASONING:
```

REDIST Reader Prompt (Candidate Selection and Verification)

```
<student prompt above, ending with "OUTPUT:\n<think>\n">
<candidate reasoning (original draft or teacher rewrite)>
</think>
<answer>
```