Off-Trajectory Reasoning: Can LRMs Collaborate on Reasoning Trajectory?

Aochong Oliver Li
Department of Computer Science
Cornell University
aochongli@cs.cornell.edu

Tanya Goyal
Department of Computer Science
Cornell University
tanyagoyal@cornell.edu

Abstract

Large Reasoning Models (LRMs) are trained to verbalize their reasoning process, yielding strong gains on complex tasks. This transparency also opens a promising direction: multiple reasoners can directly collaborate on each other's thinking within a shared trajectory, yielding better inference efficiency and exploration. A key prerequisite, however, is the ability to assess the usefulness of and build on another model's partial thinking—we call this *off-trajectory reasoning*. Our paper investigates a critical question: can standard *solo-reasoning* training pipelines deliver desired *off-trajectory* behaviors? We propose twin tests that capture the two extremes of the off-trajectory spectrum, namely **Recoverability**, which tests whether LRMs can backtrack from "distractions" induced by misleading reasoning traces, and **Guidability**, which tests their ability to build upon correct reasoning from stronger collaborators. Our study evaluates 15 open-weight LRMs (1.5B–32B) and reveals a counterintuitive finding—"stronger" LRMs on benchmarks are often more fragile under distraction. Moreover, all models tested fail to effectively leverage guiding steps from collaborators on problems beyond their inherent capabilities with solve rates remaining under 9.2%. Finally, we conduct control studies to isolate the effects of three factors in post-training on these behaviors: the choice of distillation teacher, the use of RL, and data selection strategy. Our results provide actionable insights for training natively strong reasoning collaborators; e.g., we find that suboptimal recoverability behaviors of teacher models are transferred to distilled students even if the distillation trajectories are correct. Taken together, this work lays the groundwork for evaluating multi-model collaborations in shared reasoning trajectories and highlights the limitations of off-the-shelf LRMs.

1 Introduction

LLMs with thinking abilities, such as OpenAI's o-series [23], DeepSeek-R1 [15], and Qwen3 Thinking [48], have recently emerged as the frontier models for complex reasoning tasks like mathematics and coding. These models, trained with reinforcement learning with verifiable rewards (RLVR) [43] or distillation [20], learn to verbalize their intermediate reasoning in language and exhibit self-reflective behaviors [13], such as verifying answers or seeking alternative approaches.

This transparency opens up a promising direction—stronger LRM collaborators or even human overseers can directly intervene on an LRM's ongoing reasoning and exert direct control over its thinking. This new paradigm, as demonstrated in Figure 1, can have positive implications including but not limited to:

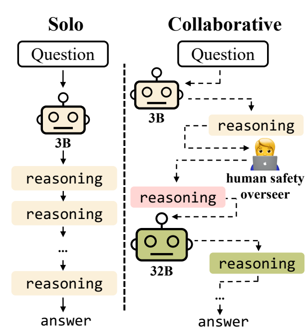

Figure 1: Comparison of solo (left) vs. collaborative reasoning (right).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

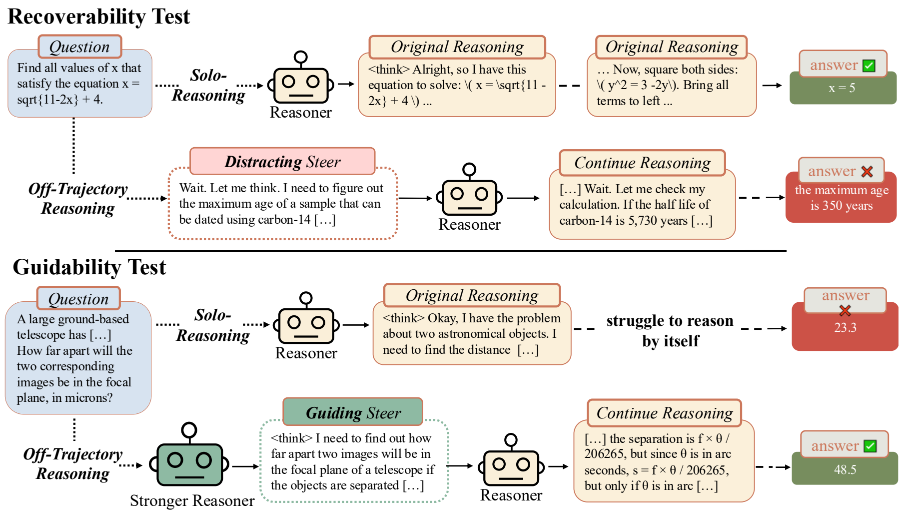

Figure 2: Illustration of the twin tests: we perturb a model's reasoning trajectories with off-trajectory steers to evaluate its *recoverability* (under a distracting steer) or *guidability* (under a guiding steer). The distracting steer is sampled from the same reasoner but for a different question.

(1) **Efficiency**: balancing performance and inference speed, large-scale LRMs should ideally focus on challenging derivations and offload routine sub-steps (e.g., arithmetic checking) to smaller models [1, 6]. (2) **Exploration**: models/humans with complementary expertise can broaden the reasoning search by spawning diverse branches [8, 39, 38] and composing their skills to solve cross-domain tasks. (3) **Safety**: an overseer model or even humans can directly intervene to steer the ongoing reasoning in a safer direction rather than abruptly terminate the reasoning process [46, 50, 28].

Most LRMs today are trained and evaluated to generate complete reasoning processes on their own, which we term *solo-reasoning*. But can they collaborate with other reasoners—models, humans, or programs—in real time within their trajectories? While some recent work has explored these possibilities [1, 6], it remains unclear whether solo-reasoning LRMs are equipped to effectively leverage partial reasoning trajectories from other collaborators due to the associated distribution shift. Ideally, LRMs should integrate useful insights from collaborators and reliably backtrack from incorrect or unhelpful inputs, even if these traces do not naturally occur in their distribution. We call this capability off-trajectory reasoning and ask: **Can solo-reasoning LRMs collaborate with off-distribution trajectories?**

We approach this question by decomposing off-trajectory reasoning into two complementary parts, **recoverability** and **guidability**, and evaluating both in simulated collaboration scenarios (see Figure 2). The recoverability test is designed to evaluate if LRMs can robustly backtrack from erroneous reasoning from collaborators to continue their original correct trajectories. At the other end of the spectrum, the guidability test evaluates if LRMs can successfully build upon correct yet incomplete reasoning from guiding models to tackle problems that are unable to solve by solo-reasoning.[1]

We systematically evaluate 15 open-weight LRMs on a suite of five math benchmarks [35, 36, 19, 32, 16]. Counterintuitively, we find that stronger reasoning models are more prone to failure under off-trajectory distractions. In the recoverability test, their performance drops to 74.9% on problems they originally solved with 100% success rate. At the same time, the guidability test reveals that LRMs fail to leverage useful hints to continue from other models' correct trajectories, even when correct answers are already present in these trajectories. Overall, our results present a sobering view into LRMs' "reasoning capabilities"—LRMs can neither reject distracting nor build upon useful off-trajectory inputs. Moreover, we show that the current practice of over-optimizing for benchmark performances do not account for broader reasoning capabilities, of which off-trajectory reasoning is an intrinsic part.

[1]We systematically test for correctness of reasoning in this paper. However, our framework can be extended for other aspects of alignment. For example, can solo LRMs robustly reject unsafe collaborator trajectories?

Next, we investigate how decisions in post-training, particularly the choice of teacher models for distillation, training data selection strategies, and RL training after distillation, impact recoverability and guidability. Through carefully designed control studies, we discover that (1) the recoverability of the teacher model directly influences the student's recoverability, despite training being limited to correct trajectories that do not exhibit recoverability errors, (2) RL can further improve both recoverability and guidability when supervised fine-tuning (SFT) saturates, and (3) aggressively reducing distillation data quantity based on quality filtering can lead to high variance in recoverability across checkpoints for similar benchmark scores.

As a step towards multi-reasoner collaboration, our work makes these key contributions:

1. We introduce the **Recoverability** and **Guidability** tests as a systematic framework for evaluating off-trajectory reasoning. Our setup complements existing standard solo-reasoning benchmarks by offering a different perspective on reasoning performance. (§2)
2. Equipped with this framework, we evaluate 15 open-weight LRMs for off-trajectory reasoning. Our analysis reveals **key limitations of "strong" solo reasoners** and shows that they consistently fail at exploiting correct guidance to improve beyond their inherent capability limits. (§3)
3. We conduct the first control studies on the **direct effects of post-training decisions**—distillation teacher models, RL fine-tuning, and data filtering—on recoverability and guidability. Our results provide actionable insights for training solo-reasoners to be robust to off-distribution distractions and to exhibit better performance in off-trajectory reasoning. (§4)

2 Twin Tests for Off-Trajectory Reasoning

Preliminaries and Notation. Let M be a reasoning model and (q, a^*) be a training or test data point. In standard solo-reasoning, M generates a reasoning trajectory $\mathbf{r} = [r_1, r_2, \ldots, r_k]$ and a final answer a for an input question q, i.e., $(\mathbf{r}, a) \sim M(\cdot \mid q)$. We use r_i to refer to a *reasoning unit*, the granularity of which can be flexibly determined.

In contrast, in the collaborative setting, multiple models or different instantiations of the same model contribute different parts to the reasoning trajectory $\mathbf{r}$. Recent work has explored some collaboration strategies, such as dynamically off-loading reasoning sub-parts to weaker/stronger models [47, 52, 1], tooling [27] or aggregating parallel samples [51, 39] during both training and inference.

The success of such collaboration hinges on the main model M's ability to process and build upon a trajectory mixing both in- and off-distribution reasoning units $\mathbf{r} = [r^M, r^{M'}, r^{M''}, \ldots, r^{M'''}]$. In this paper, we instantiate a simplified setup of two-model collaboration to probe off-trajectory reasoning capabilities in frontier open-weight LRMs.

Two-Model Setup We simulate a collaboration between two reasoning systems, where the main model M and the collaborator M_{steer} jointly contribute to an off-trajectory reasoning $[r^{\text{og}}, r^{\text{steer}}]$. In practice, we construct r^{og} by sampling from the main model M and stopping generation at m tokens, i.e., $|r^{\text{og}}| = m$. Similarly, r^{steer} is sampled from the collaborator with $|r^{\text{steer}}|$ limited to n tokens. To measure off-trajectory reasoning performance, we concatenate these two incomplete trajectories to construct a shared off-distribution trajectory. Finally, we sample a reasoning completion and final answer from M conditioned on the original question and this trajectory.

$$(\mathbf{r}^{\text{off}}, a^{\text{off}}) \sim M(\cdot \mid q, [r^{\text{og}}, r^{\text{steer}}])$$

For domains with verifiable rewards, we can measure the success of this off-trajectory completion by computing the accuracy of the final answer, i.e., $\mathbb{E}_{(q,a^*)\sim\mathcal{D}}\left[\mathbb{1}\{\, a^{\text{off}} = a^* \,\}\right]$

Considerations for designing the steer. This simplified setup allows us to flexibly simulate the two extreme effects r^{steer} can have on the main model M. At one end, the steer can be *distracting*: it misleads M away from its original correct trajectory and steers it down an incorrect path. At the other end, the steer can have *guiding* effects: it provides hints that can potentially guide M towards a correct solution for challenging problems beyond its capability boundaries.

Based on these desiderata, we design twin tests: (i) **Recoverability**, which tests whether LRMs can resist a distracting steer and backtrack to previous reasoning, and (ii) **Guidability**, which tests models' abilities to successfully leverage a guiding steer to surpass their solo-reasoning ability.

These twin tests differ mainly in two aspects: the selection of test questions q and the construction of steered trajectories $[r^{\text{og}}, r^{\text{steer}}]$. Given an original test set $\mathcal{D}$ and test model M, our protocol automatically instantiates an M-specific off-trajectory dataset for both tests separately, i.e., $\mathcal{D}_M^{\text{test}} = \{(q, [r^{\text{og}}, r^{\text{steer}}], a^*)\}$. The overall process for this is shown in Figure 2 and described below.

2.1 Recoverability Test

Selecting test data points $\{(q, a^*)\}$. Our goal is to test how well M can backtrack from a distracting steer and still output the correct answer a^*. For a given test model M, we select the subset of test questions that M can correctly answer in solo-reasoning, i.e., $a = a^*$, where $(r, a) \sim M(\cdot \mid q)$. This selection can isolate the effects of distracting steers from M's inherent capabilities .

Constructing steered trajectories. The trajectory consists of two parts: r^{og} and r^{steer}. We truncate r, the reasoning trajectory from solo-reasoning, to the first m tokens to obtain r^{og}. In our experiments, described in § 3.1, we vary m as a fraction of the total number of tokens in r.

We require r^{steer} to be a strong distractor for the test model M. However, it is difficult to determine *a priori* which model M_{steer} and steer r^{steer} will achieve this reliably. Therefore, we simulate the distraction r^{steer} by sampling from M itself, but conditioned on a different question q'. So, if M is distracted to blindly complete r^{steer}, its reasoning is then guaranteed to be incorrect. In practice, we control the length of r^{steer} by truncating it to the first n tokens of r', where $(r', a') \sim M(\cdot \mid q')$. In our experiments, we control the strength of the distractor by varying n (i.e., $|r^{\text{steer}}|$) and the insertion point by varying m (i.e., $|r^{\text{og}}|$). Exact experiment details are provided in § 3.1.

2.2 Guidability Test

Selecting test data points $\{(q, a^*)\}$. In the guidability test, we aim to study whether M can effectively leverage a *guiding steer*, i.e., a correct partial reasoning, for questions it struggles with during solo-reasoning. Therefore, we select the subset of test questions for which the solo-reasoning solve rate is either 0 or 1 out of 8 samples.

Constructing steered trajectories. First, unlike the recoverability test, we do not include M's own reasoning trace r^{og} in steered trajectory (i.e., set $m = 0$). This is because r^{og} might already contain errors that anchor M in the wrong direction, thereby confounding the measurement of guidability.

We construct r^{steer} using a stronger reasoner M_{steer} as the guide, i.e., with a higher benchmark performance than M. Figure 3 illustrates this. To test whether M can build on M_{steer}'s correct reasoning, we only provide the first n tokens of the complete trajectory. In practice, we vary the "amount" of guidance by varying n to different fractions of the complete trajectory from the guide. Moreover, we use multiple guiding models to construct independent steers for each q. This allows us to measure guidability under different off-trajectory distributions and amount of guidance.

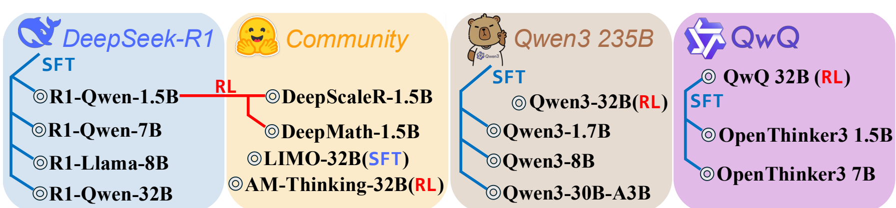

Figure 3: 15 open-weight LRMs grouped into four families. The branches indicate the source from which LRMs are derived, and the colors indicate SFT/RL training methods.

3 Off-the-shelf Evaluation & Results

3.1 Experiment Setup

Datasets and Benchmarks. We run our experiments on 15 open-weight models. To illustrate the relationships between these LRMs, we group them into four families (see Figure 3):

- **DeepSeek-R1** [15]: `R1-Qwen-1.5B/7B/32B` and `R1-Llama-8B` are distilled from `DeepSeek-R1` using supervised fine-tuning (SFT).

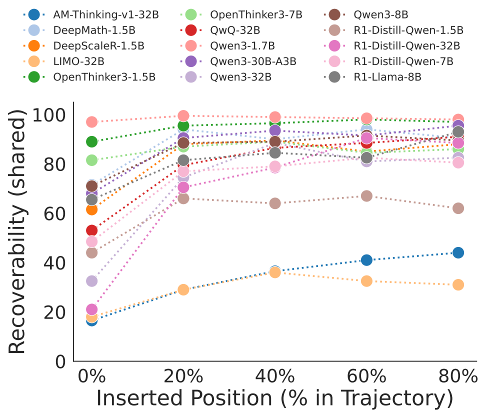

Figure 4: Recoverability (shared) across positions (%) of the original trajectory for 15 LRMs

Model	**Teach.** (%)	**Ans.?** (%)	Δ
R1-Qwen-1.5B	28.4	25.6	2.8
DeepScaleR-1.5B	29.8	23.3	6.5
R1-Llama-8B	35.0	21.8	13.2
DeepMath-1.5B	27.1	22.9	4.2
OpenThinker3-1.5B	32.7	26.9	5.8
Qwen3-1.7B	29.9	18.0	11.9
R1-Qwen-7B	19.7	12.1	7.6
LIMO-32B	21.5	10.2	11.3
OpenThinker3-7B	20.6	13.8	6.8
R1-Qwen-32B	22.5	11.2	11.3
Avg.	26.7	18.6	8.1

Table 2: Analysis of guidability results. Teach. = guidability score (individual); Ans.? = fraction of steers already containing the correct answer; Δ = Teach. − Ans. (pp).

- **Qwen3** [48]: `Qwen3-32B` is directly trained with RL for reasoning without distillation, while `Qwen3-1.7B/8B/30B-A3B` are distilled from `Qwen3-235B` and `Qwen3-32B`.
- **QwQ**: `QwQ-32B` [40] is directly trained with RL from the `Qwen2.5-32B-Base` model to enhance its reasoning capabilities. `OpenThinker3-1.5B/7B` [14] are based on `Qwen2.5-Instruct` and distilled from `QwQ-32B` on 1.2M curated math and coding examples.
- **Community**: `DeepScaleR-1.5B` [34] and `DeepMath-1.5B` [18] are trained with RL on `R1-Qwen-1.5B` using DeepScaleR and DeepMath datasets, respectively. `LIMO-32B` [49] is SFT from `Qwen2.5-32B-Instruct` on the LIMO dataset of 817 examples. Finally, `AM-Thinking-32B` [25] is a `Qwen2.5-32B-Base` model first distilled on 2.84M examples, and then trained with RL on 54K math and coding questions.

We evaluate on a pool of 1,507 math questions sourced from five standard benchmarks, AIME-2024 [35], AIME-2025 [36], MATH-500 [19], Minerva (math subset) [32], and OlympiadBench [16].

Hyperparameter Settings. All LRMs are evaluated under the same hyperparameter settings: maximum tokens of 32K, temperature 0.6, top-p 0.95, and no system prompt. For each question, we sample 8 completions and report the average Pass@1 over samples.

Recoverability and Guidability Setup. Following the protocols in §2.1, we sample 200 original trajectories r^{og} and 50 trajectories as distracting steers r^{steer} for each LRM. By default, we set n, i.e., $|r^{\text{steer}}|$ to be 0.2 times the length of the full *distracting* trajectory; this leaves sufficient tokens for *off-trajectory* completion. We set m, i.e., $|r^{\text{og}}|$, to be 0, 0.2, 0.4, 0.6, and 0.8 times the length of the original reasoning from the main model. We report recoverability on two subsets: (1) *shared* subset that includes questions that all 15 LRMs can fully solve (8 out of 8), and (2) *individual* subset that samples questions independently for each LRM following the criterion defined in § 2.

We instantiate the guidability tests using `DeepSeek-R1`, `Qwen3-235B`, and `QwQ-32B` as M_{steer} to sample *guiding* steers r^{steer}. Since the best 5 LRMs almost saturate the benchmarks, we only evaluate on the remaining 10 LRMs that have enough questions with solve rate $\leq \frac{1}{8}$ (Table 9). We set n, i.e., $|r^{\text{steer}}|$, to be 0.2, 0.4, 0.6 and 0.8 times the total tokens in the guide's reasoning. Similar to the recoverability test, we report guidability scores on two subsets: *shared* (intersection across the 10 evaluated models) and *individual* (per model).

3.2 Results

Our main results are shown in Table 1. We group models into low, medium, and high tiers based on their solo-reasoning performance (reported in the *Avg. Benchmark* column) and report recoverability and guidability results on both shared and individual subsets.

Finding 1: Stronger solo-reasoners $\neq$ stronger collaborators. Surprisingly, we find that recoverability and guidability are largely orthogonal to LRMs' solo-reasoning performance. Particularly, we highlight models in the *low* benchmark tier such as `OpenThinker3-1.5B` and `Qwen3-1.7B` that exhibit substantially better recoverability than *medium* and *high* tier models like `QwQ-32B` and `Qwen3-32B`. Noticeably, the best performing solo-reasoning model `AM-Thinking-32B` reports the

Model	Family	Benchmark Avg.	Recoverability Sh.	Recoverability Ind.	Guidability Sh.	Guidability Ind.
Low Benchmark Scores						
R1-Qwen-1.5B	DS-R1	47.5	$60.6_{\uparrow+2}$	$38.6_{\uparrow+2}$	$3.0_{\uparrow+0}$	$28.4_{\uparrow+5}$
DeepScaleR-1.5B	Comm.	53.3	$82.4_{\uparrow+7}$	$52.9_{\uparrow+5}$	$4.1_{\uparrow+1}$	$29.8_{\uparrow+5}$
R1-Llama-8B	DS-R1	54.1	$81.4_{\uparrow+5}$	$49.6_{\uparrow+3}$	$8.7_{\uparrow+4}$	$35.0_{\uparrow+7}$
DeepMath-1.5B	Comm.	54.8	$88.0_{\uparrow+9}$	$61.8_{\uparrow+6}$	$3.4_{\downarrow-2}$	$27.1_{\uparrow+1}$
OpenThinker3-1.5B	QwQ	59.2	$95.2_{\uparrow+9}$	$71.8_{\uparrow+8}$	$5.7_{\downarrow-1}$	$32.7_{\uparrow+4}$
Qwen3-1.7B	Qwen3	59.9	$98.4_{\uparrow+9}$	$74.6_{\uparrow+9}$	$6.1_{\uparrow+0}$	$29.9_{\uparrow+2}$
Medium Benchmark Scores						
R1-Qwen-7B	DS-R1	64.6	$73.5_{\downarrow-1}$	$45.8_{\downarrow-2}$	$6.0_{\downarrow-2}$	$19.7_{\downarrow-6}$
LIMO-32B	Comm.	67.3	$29.3_{\downarrow-7}$	$18.5_{\downarrow-7}$	$8.8_{\uparrow+0}$	$21.5_{\downarrow-5}$
OpenThinker3-7B	QwQ	72.1	$85.6_{\uparrow+1}$	$74.5_{\uparrow+5}$	$9.1_{\uparrow+0}$	$20.6_{\downarrow-7}$
R1-Qwen-32B	DS-R1	72.3	$69.8_{\downarrow-6}$	$45.6_{\downarrow-6}$	$9.2_{\uparrow+0}$	$22.5_{\downarrow-6}$
High Benchmark Scores						
Qwen3-8B	Qwen3	79.1	$85.9_{\uparrow+0}$	$68.8_{\uparrow+1}$	N/A	N/A
QwQ-32B	QwQ	80.5	$79.7_{\downarrow-5}$	$62.6_{\downarrow-1}$	N/A	N/A
Qwen3-32B	Qwen3	81.0	$71.8_{\downarrow-8}$	$56.9_{\downarrow-5}$	N/A	N/A
Qwen3-30B-A3B	Qwen3	81.1	$87.8_{\downarrow-2}$	$60.0_{\downarrow-5}$	N/A	N/A
AM-Thinking-32B	Comm.	82.6	$33.4_{\downarrow-13}$	$25.3_{\downarrow-13}$	N/A	N/A

Table 1: **Results for 15 LRMs from four families.** Columns report benchmark averages and recoverability/guidability scores for *shared* (Sh.) and *individual* (Ind.) subsets. Models are grouped into low/medium/high tiers by *Benchmark Avg*. Subscripts indicate rank changes relative to the benchmark ranking ($+k$ rise, $-k$ drop); green (↑) denotes improvement, red (↓) decline. "DS-R1" = DeepSeek-R1 family, "Comm." = Community models. N/A = not evaluated. Our results show that the benchmark performances are largely orthogonal to recoverability.

second worst recoverability performance. Similarly, `LIMO-32B`—claimed to surpass prior SFT approaches using only 1% of training data—only recovers less than 30% of the time. Across models, we observe an average of 25.1% degradation in their reasoning capabilities, when their trajectories are perturbed with tangential distractions.

In addition, our results show that all LRMs report exceptionally low guidability scores; none of the models report $>10\%$ on the shared subset. Taken together, these findings suggest that **models optimized heavily for popular benchmarks may have hidden vulnerabilities, particularly in off-trajectory reasoning**. Our twin tests successfully surface such limitations.

Finding 2: The beginning of model reasoning is critical for recovery. To better understand the recoverability trends in Table 1, we visualize the recovery rates separately for different percentages (%) of the original thinking trajectory where the distracting steer is inserted. Figure 4 shows these results.[2] Interestingly, we observe a consistent pattern across models—distraction at the very start (0%) of the trajectory leads to the largest degradation. This is surprising as models typically only restate the question in the opening and rarely include actual problem solving. Given these results, we hypothesize that restating the question at the start is critical for models to anchor later reasoning.

To test our hypothesis, we conduct an ablation that re-instantiates the recoverability tests but preserves the first paragraph of the original trajectory. We find that most LRMs exhibit noticeable improvements across positions after this change, especially at the 0% position[3]. In fact, the average recoverability score exceeds 83.5% for all models (except `LIMO-32B` and `AM-Thinking-32B`) with this small tweak in their reasoning trajectories. This clearly shows that **while restating the question does not add new information, it is critical for LRM off-trajectory reasoning**.

Finding 3: LRMs fail to leverage correct guidance to surpass their inherent limits. As Table 1 shows, all models, regardless of their solo-reasoning capabilities, struggle to effectively build upon guiding trajectories. Crucially, we find that the performance does not improve even when models are paired with their own distillation teacher, i.e., the model whose samples they were trained on (see

[2]The full set of results for both shared and individual metrics are reported in Tables 5 and 7 in the Appendix.

[3]The complete set of results is included in Table 6 in the Appendix

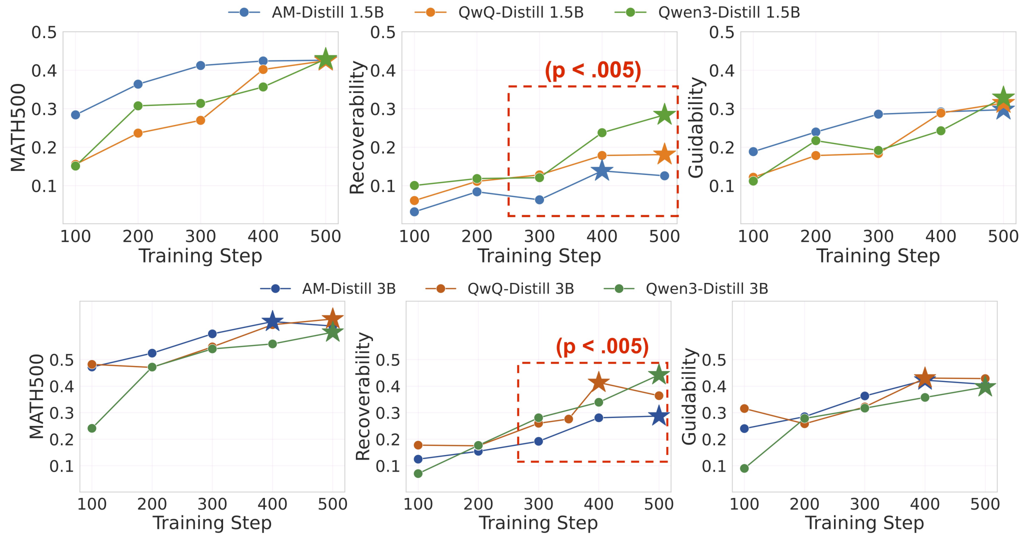

Figure 5: Qwen2.5 models (1.5B and 3B) distilled from `AM(-Thinking)-32B` show consistently lower recoverability than those distilled from `QwQ-32B` or `Qwen3-32B`, while having similar performance on benchmark and guidability; the gap is significant after step 300 ($p \leq 0.005$). Stars mark each model's peak over training steps.

Table 12 for full set of results). For example, `Qwen3-1.7B` shows no guidability gains when guided by `Qwen3-235B` compared to other models.

Further investigation reveals that **even these low guidability scores are artificially inflated.** Since we truncate the guiding steer at different lengths, it is possible that some partial r^{steer} already contain the correct answer derivation. In such cases, we expect the guidability test to be trivially easy.

In Table 2, we report the percentage of guiding steers that already contain the correct answer (Ans.? column). We find that this is true for 18.6% of steers on average (see Table 10 for breakdown by steer length). However, we find that LRMs can often fail to recognize such correct reasoning, reject the given answer and pivot to an incorrect path, resulting in the low guidability scores. This suggests that conditioning LRMs on correct but out-of-distribution traces does not enable them to successfully leverage these guiding traces and surpass their inherent capability limits.

4 Control Studies on Post-training Decisions

Section 3 shows that different LRMs exhibit distinct off-trajectory behaviors. However, these LRMs are trained on different data and derived from different base models; therefore, it remains unclear what factors in the post-training procedures drive these differences. To understand this, we conduct controlled experiments to isolate the effects of (1) teacher models used for distillation in § 4.1, (2) RL training after SFT in § 4.2, and (3) quality heuristics for data filtering in § 4.3.

4.1 How Do Teachers' Behaviors Affect Distilled Models?

Hypothesis. We observe from Table 1 that LRMs distilled from `DeepSeek-R1` generally have lower recoverability scores compared to those from `QwQ` and `Qwen3`. This is despite the fact that most of them are trained from similar base models using distillation. Therefore, we ask: *Do distilled models inherit the vulnerabilities of their teachers' off-trajectory behaviors through distillation?*

Setup. We conduct controlled experiments with three LRMs as the distillation teacher models: `AM-Thinking-32B`, `QwQ-32B`, `Qwen3-32B`. We choose the AM model since it has similar benchmark performance but significantly lower recoverability compared to QwQ and Qwen3 models in Table 1. We perform SFT on two Qwen2.5 models (1.5B and 3B) with correct trajectories from each teacher separately (more details in Appendix F).

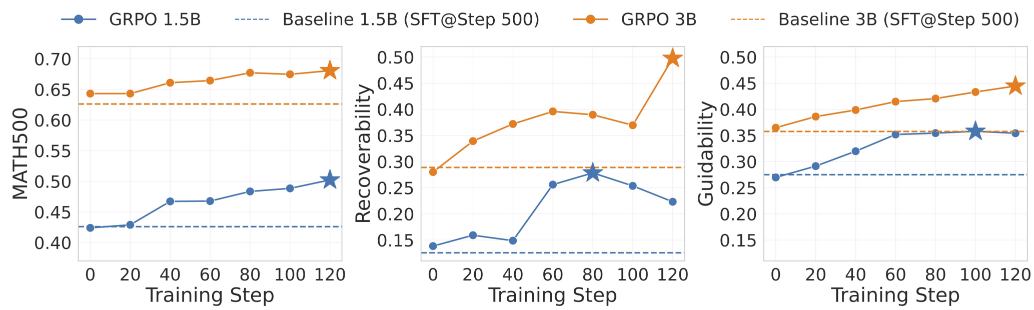

Figure 6: GRPO 1.5B and 3B (from SFT@Step 400) show noticeable gains on benchmark, recoverability, and guidability compared to the initial checkpoint and baselines (SFT@Step 500). This improvement is consistent over RL training. Stars mark the peak values over training steps.

We evaluate the distilled models (AM-/QwQ-/Qwen3-Distill 1.5B/3B) on MATH-500 for benchmark performance and twin tests. Figure 5 reports the results and highlights checkpoints with significant differences ($p \leq 0.005$) based on two-sample t-tests.

Results: Students mirror their teacher's recoverability performance. Our results show that AM-Distill models show significantly lower recoverability than QwQ- and Qwen3-Distill counterparts after step 300, despite similar benchmark and guidability scores. This recoverability gap persists across all model sizes that we tested and also remains consistent at different positions of the reasoning trajectories (Appendix F).

Our results highlight that correctness should not be the sole criterion for selecting teacher trajectories. Instead, other vulnerabilities of the teacher model should be accounted for as these may be distilled into student models. Our twin tests provide a useful criterion for selecting teachers, and can be combined with other metrics of selection.

4.2 Can RL Further Improve Off-Trajectory Reasoning after SFT Saturates?

Hypothesis. In Table 1, we do not observe a consistent advantage of RL over SFT distillation on twin tests. However, training recipes of these models are different, making it impossible to draw concrete conclusions about RL's impact. Here, we ask: *Can RL further improve both recoverability and guidability even after SFT has saturated?*

Setup. We use distillation checkpoints from Section 4.1—AM-Distill 1.5B and 3B models at step 400—as the initial policy for RL training. This choice is motivated by: (1) we observe that SFT saturates on benchmarks and twin tests after step 400; and (2) AM-Distill is shown to perform poorly in recoverability, making it more suitable to test the effects of RL. We train both models on the MATH8K dataset with Grouped Relative Policy Optimization (GRPO) [43].

Results: RL training reports massive improvements in recoverability. Figure 6 shows the impact of RL training on benchmark scores, recoverability and guidability. While all scores improve with RL, we see a noticeably high recoverability improvement (e.g., 15.3%-28.9%) accompanying a slight increase in benchmark scores (5.4%-7.6%) and guidability (8.3%-8.7%). Notably, RL training completely bridges the gap in recoverability that we observed in Figure 5 between AM-Distill and QwQ-/Qwen3-Distill models. We hypothesize that outcome-based RL improves recoverability by exposing models to noisy trajectories and explicitly rewarding successful recoveries. In contrast, SFT training is mostly on successful demonstrations. We leave a more thorough investigation of the mechanisms behind the observed improvement to future work.

4.3 Does Less Data Always Lead to Poorer Recoverability?

Hypothesis. Recent works have shown that data quality is critical for strong reasoning capabilities [11, 2, 14]. The "Less-Is-More" (LIMO) hypothesis [49] pushes for an extreme version of this claim—a minimal amount of "high-quality" data is sufficient to elicit complicated reasoning. [49] curate the LIMO dataset of 817 examples filtered based on heuristics and support their claim with the performance of the `LIMO-32B` model on popular reasoning benchmarks. Their results imply that data quantity is less important for training LRM reasoning as long as the data quality is "high" based on their criteria. However, we observe a contrary result in Table 1 where the `LIMO-32B` model reports the worst recoverability despite decent solo-reasoning performance. To understand this, we ask: *Is the less-is-more paradigm inherently limited for off-trajectory reasoning?*

Setup. We train `Qwen2.5-3B-Base` models on two larger datasets of mixed "quality" and two smaller ones of only "high-quality" data: (1) **FULL-8K**: MATH8K dataset distilled from `QwQ-32B` in §4.1 (i.e., the same dataset used to train QwQ-Distill 3B in §4.1); (2) **FULL-8.8K**: a mix of FULL-8K and the LIMO dataset [49]; (3) **LIMO-800**: the LIMO dataset; and (4) **LIMO-600**: 600 "challenging" examples we extracted from FULL-8K, following the "LIMO" principle, i.e., classified as Level-5 difficulty and with long reasoning trajectories. We train each model with SFT until its benchmark performance plateaus. Figure 7 plots recoverability scores against benchmark scores at different checkpoints during training.

Results: To our surprise, **models trained on less data are not necessarily worse on recoverability but exhibit extremely high variance between checkpoints.** LIMO-600 and LIMO-800 3B models show markedly different levels of recoverability against similar benchmark scores. On the other hand, FULL-8K and FULL-8.8K models trained on larger datasets have minimal variance across checkpoints with the same benchmark scores.

Our results show that "over-optimizing" benchmarks through aggressive data filtering could introduce unwanted biases in off-trajectory behaviors that are not captured by standard solo-reasoning evaluations. In addition, our tests can complement existing criteria for selecting checkpoints with higher robustness to out-of-distribution scenarios.

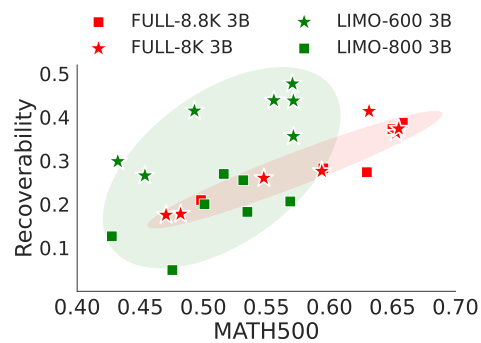

Figure 7: LIMO-600/-800 3B models exhibit greater variance in recoverability than FULL-8K/8.8K 3B. Colors: FULL, LIMO. Markers: square = contains data from LIMO-800, star = otherwise. We observe that model checkpoints trained on high-quality but limited quantity of data show high variance in recoverability scores across similar benchmark score values.

5 Related Work

Large Reasoning Models. Recent post-training advances have led to massive improvements on math and coding benchmarks [22, 15], as exhibited by both closed- and open-source LRMs since the release of OpenAI's o-1 [23], e.g., [15, 48, 14, 49, 25]. These models are typically trained to produce extended reasoning traces using RL algorithms such as Proximal Policy Optimization (PPO) [42], Grouped Relative Policy Optimization (GRPO), and related variants [43], typically with verifiable rewards. At smaller scales (under 32B parameters), reasoning models like R1-Qwen-Distill series [15] and Qwen3 family [48] are primarily trained with distillation [20]. Additionally, the open-source community has also released artifacts that further train these models with RL. In our study, we analyze 15 representative open-weight LRMs spanning diverse model families and training paradigms.

LRM Reasoning Intervention and Collaboration. Recent studies intervene on LRM reasoning process to understand and control their behaviors, including perturbing intermediate steps to examine their faithfulness [3, 4], improve instruction following and alignment behaviors [46], or interpret [31, 37] and stress-test cognitive behaviors [13]. [45] examine the impact of thinking patterns on outcome correctness, while [17, 31] systematically categorize different types of reasoning strategies and errors. In addition, our work also sits within the prior work on teacher–student framework for augmenting model reasoning [21, 1, 5]. In a closely related work, [17] investigates LRMs' ability to recover from unhelpful thoughts. Our twin tests also intervene on reasoning but differ in their goal of simulating extreme scenarios of multi-model collaboration.

Our work is also closely related to hybrid parallel and serialized scaling approaches [38], including offloading challenging reasoning parts to larger models [1] and orchestrating different models for high-level planning and downstream execution [30]. Our work evaluates how solo-reasoning LRMs can fail when routed onto a shared reasoning trajectory.

6 Limitations & Future Work

Our study conducts an initial systematic investigation into the fragility of LRM off-trajectory reasoning. In this work, we report the results of the Recoverability and Guidability twin tests on math reasoning benchmarks, reflecting that most open-weight LRMs are primarily post-trained on math datasets. Our framework, however, can be straightforwardly extended to other domains. We encourage future work to extend our framework to other domains, such as coding [24, 26, 7], science [44, 41, 12], and logical reasoning tasks [10, 33, 9].

For better control, our experiments use a two-model, single-turn simulation setting. However, real-world multi-agent, multi-turn interactions can be more complex; we view this work as laying the foundation for studying richer collaborative dynamics. Additionally, we make certain design decisions in our twin tests that can be studied further. For instance, in Recoverability, distractors are sampled from the same model on a different question to model the "distracting effects" of erroneous traces. This choice may make distractors stylistically and syntactically similar to the original reasoning, potentially overstating the brittleness of LRMs relative to distractors from other models.

7 Conclusion

In this work, we investigate off-trajectory reasoning in LRMs—their ability to "think" on trajectories steered by other reasoners. We introduce Recoverability and Guidability tests to evaluate model robustness under off-trajectory reasoning, which test (i) the ability to backtrack to original correct trajectories conditioned on distracting steers, and (ii) the ability to effectively use guidance from off-distribution traces. Our evaluation of 15 open-weight LRMs on both tests reveals that all open-weight LRMs perform poorly on these tests, highlighting limitations of standard solo-reasoners in collaborative settings. Finally, control studies show that recoverability is directly shaped by distillation teachers, can be improved with RL fine-tuning, and becomes more unpredictable as the size of the distillation dataset shrinks. These results offer valuable insights for future work to advance collaborative reasoning systems.

8 Acknowledgements

This project was partially supported by NSF grant IIS-2433072, and a gift from Google. We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc, with support from Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation.

References

[1] Yash Akhauri, Anthony Fei, Chi-Chih Chang, Ahmed F AbouElhamayed, Yueying Li, and Mohamed S Abdelfattah. Splitreason: Learning to offload reasoning. *arXiv preprint arXiv:2504.16379*, 2025.

[2] Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, et al. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025.

[3] 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.

[4] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[5] Edoardo Cetin, Tianyu Zhao, and Yujin Tang. Reinforcement learning teachers of test time scaling. *arXiv preprint arXiv:2506.08388*, 2025.

[6] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[8] Mouxiang Chen, Binyuan Hui, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Jianling Sun, Junyang Lin, and Zhongxin Liu. Parallel scaling law for language models. *arXiv preprint arXiv:2505.10475*, 2025.

[9] Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

[10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[11] Quy-Anh Dang and Chris Ngo. Reinforcement learning for reasoning in small llms: What works and what doesn't. *arXiv preprint arXiv:2503.16219*, 2025.

[12] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[13] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[14] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL `https://aclanthology.org/2024.acl-long.211/`.

[17] Yancheng He, Shilong Li, Jiaheng Liu, Weixun Wang, Xingyuan Bu, Ge Zhang, Zhongyuan Peng, Zhaoxiang Zhang, Zhicheng Zheng, Wenbo Su, et al. Can large language models detect errors in long chain-of-thought reasoning? *arXiv preprint arXiv:2502.19361*, 2025.

[18] Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025.

[19] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=7Bywt2mQsCe`.

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

[21] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL `https://aclanthology.org/2023.acl-long.830/`.

[22] Yichen Huang and Lin F Yang. Gemini 2.5 pro capable of winning gold at imo 2025. *arXiv preprint arXiv:2507.15855*, 2025.

[23] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[24] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

[25] Yunjie Ji, Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Han Zhao, and Xiangang Li. Am-thinking-v1: Advancing the frontier of reasoning at 32b scale. *arXiv preprint arXiv:2505.08311*, 2025.

[26] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[27] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[28] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

[29] Hynek Kydlíček. Math-Verify: Math Verification Library. URL `https://github.com/huggingface/math-verify`.

[30] Byeongchan Lee, Jonghoon Lee, Dongyoung Kim, Jaehyung Kim, and Jinwoo Shin. Collaborative llm inference via planning for efficient reasoning. *arXiv preprint arXiv:2506.11578*, 2025.

[31] Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, et al. The cot encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think. *arXiv preprint arXiv:2505.10185*, 2025.

[32] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[33] Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

[34] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

[35] MAA. American Invitational Mathematics Examination 2024 I & II. `https://maa.org/maa-invitational-competitions/`, 2024.

[36] MAA. American Invitational Mathematics Examination 2025 I & II. `https://maa.org/maa-invitational-competitions/`, 2025.

[37] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, et al. Deepseek-r1 thoughtology: Let's think about llm reasoning. *arXiv preprint arXiv:2504.07128*, 2025.

[38] Jiayi Pan, Xiuyu Li, Long Lian, Charlie Snell, Yifei Zhou, Adam Yala, Trevor Darrell, Kurt Keutzer, and Alane Suhr. Learning adaptive parallel reasoning with language models. *arXiv preprint arXiv:2504.15466*, 2025.

[39] Jianing Qi, Xi Ye, Hao Tang, Zhigang Zhu, and Eunsol Choi. Learning to reason across parallel samples for llm reasoning. *arXiv preprint arXiv:2506.09014*, 2025.

[40] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL `https://qwenlm.github.io/blog/qwq-32b/`.

[41] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[43] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[44] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[45] Pengcheng Wen, Jiaming Ji, Chi-Min Chan, Juntao Dai, Donghai Hong, Yaodong Yang, Sirui Han, and Yike Guo. Thinkpatterns-21k: A systematic study on the impact of thinking patterns in llms. *arXiv preprint arXiv:2503.12918*, 2025.

[46] Tong Wu, Chong Xiang, Jiachen T Wang, G Edward Suh, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention. *arXiv preprint arXiv:2503.24370*, 2025.

[47] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

[48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[49] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=T2TZ0RY4Zk`.

[50] Yichi Zhang, Zihao Zeng, Dongbai Li, Yao Huang, Zhijie Deng, and Yinpeng Dong. Realsafe-r1: Safety-aligned deepseek-r1 without compromising reasoning capability. *arXiv preprint arXiv:2504.10081*, 2025.

[51] Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. The majority is not always right: Rl training for solution aggregation. *arXiv preprint arXiv:2509.06870*, 2025.

[52] Ruiyang Zhou, Shuozhe Li, Amy Zhang, and Liu Leqi. Expo: Unlocking hard reasoning with self-explanation-guided reinforcement learning. *arXiv preprint arXiv:2507.02834*, 2025.

A Large Language Model Usage

In this paper, we use AI with great caution for polishing the language of some texts that are originally written by the authors.

B LLM-as-a-judge Prompt

```
### System Prompt
You are an unbiased examiner who evaluates whether a student's answer to a
given question is correct.
Your task is to determine if the student's final answer matches the
standard answer provided, based solely on correctness and the question's
specific requirements.
Do not perform any additional calculations or reinterpret the question.
Simply compare the student's answer to the standard answer to determine if
it satisfies the question's requirements.

Focus strictly on:
1.  Understanding the exact requirement of the question.
2.  Comparing the student's final answer directly and rigorously to the
provided standard answer.
3.  Your task is not to solve the problem but to determine whether the
student's answer is correct based on the question's requirements.  Avoid
any unnecessary analysis, assumptions, or re-solving the problem.

Note:
- For intervals/ranges:  The student's answer must cover the EXACT SAME
range as the standard answer, NOT just any single value or subset within
that range;
- If the standard answer contains multiple solutions connected by
'or'/'and', all of them must be listed in the student's answer;
- If student's response does not mention any answer, it is considered
WRONG;
- You must be deterministic and rigorous - always declare the answer as
either CORRECT or WRONG;
- Small rounding differences are permitted if all the derivation steps are
correct.

Your response must include:
### Short Analysis
Provide a short and evidence-backed analysis between <analysis> </analysis>
tags, in which you should extract the final solution value from the
standard answer and the student's answer and judge whether they are the
same.

### Correctness
Based on the analysis, you should report a label CORRECT or WRONG
between <judge> </judge> tags (e.g., <judge>CORRECT</judge> or
<judge>WRONG</judge>).

### User Prompt
Problem:  {problem}

Standard Answer:  {standard_answer}

Student Answer:  {student_answer}
```

Table 3: LLM-as-a-judge prompt template for evaluating model responses

To ensure accurate scoring for evaluations in §3, we first validate all responses with `Math-Verify` [29] and double check with `DeepSeek-V3` as a judge. We prompt DeepSeek-V3 for responses that are labeled as wrong by math-verify. Table 3 contains the exact prompt.

C Benchmark Results

Here, we provide all 15 LRM performance on five math benchmarks. The *Avg.* column is the same as the one in Table 1.

Model	**AIME 24**	**AIME 25**	**MATH-500**	**Minerva**	**Olympiad**	**Avg.**
Low Benchmark Scores						
R1-Qwen-1.5B	30.4	21.7	84.2	47.6	53.7	47.5
R1-Llama-8B	42.9	27.1	88.3	49.0	63.5	54.1
DeepMath-1.5B	37.5	29.2	90.1	54.8	62.6	54.8
DeepScaleR-1.5B	40.0	30.0	89.9	54.7	61.8	55.3
OpenThinker3-1.5B	52.1	39.6	92.2	43.7	68.4	59.2
Qwen3-1.7B	44.2	36.7	92.1	59.5	67.3	59.9
Medium Benchmark Scores						
R1-Qwen-7B	55.4	38.3	94.3	64.3	70.8	64.6
LIMO-32B	55.8	41.7	95.4	70.5	73.0	67.3
OpenThinker3-7B	63.3	58.3	96.4	64.6	77.8	72.1
R1-Qwen-32B	67.9	52.1	95.4	69.9	76.5	72.3
High Benchmark Scores						
Qwen3-8B	76.3	70.4	97.3	72.2	79.6	79.1
QwQ-32B	79.6	69.6	97.9	72.6	83.1	80.5
Qwen3-32B	78.3	71.7	97.5	75.0	82.3	81.0
Qwen3-30B-A3B	77.5	73.8	97.6	74.1	82.2	81.1
AM-Thinking-32B	80.4	77.9	98.4	72.8	83.5	82.6

Table 4: **Benchmark performance** (%) of 15 thinking LRMs. "Olympiad" stands for OlympiadBench and "Minerva" is the math subset in Minerva benchmark. "Avg" = unweighted mean of AIME 24, AIME 25, MATH-500, Minerva, and OlympiadBench.

D Recoverability Test

Table 5 reports a breakdown of model recoverability performance on shared subset across different positions (%) of the original trajectories. Table 6 reports the results of ablation study explained in §3.2, where the first paragraph of model reasoning is preserved. The subscripts in Table 6 equals the difference between the major numbers in Table minus the corresponding numbers in Table 5 to show the changes in recoverability induced by the small tweak in trajectory.

Model	0%	20%	40%	60%	80%	Avg.	Benchmark Avg.
R1-Distill-Qwen-1.5B	44.0	66.0	64.0	67.0	62.0	60.6	47.5
R1-Llama-8B	65.5	81.5	84.5	82.5	93.0	81.4	54.1
DeepMath-1.5B	71.5	94.0	90.0	94.0	90.5	88.0	54.8
DeepScaleR-1.5B	61.5	88.0	89.5	85.0	88.0	82.4	53.3
OpenThinker3-1.5B	89.0	95.5	96.5	98.0	97.0	95.2	59.2
Qwen3-1.7B	97.0	99.5	99.0	98.5	98.0	98.4	59.9
R1-Distill-Qwen-7B	48.5	77.0	79.0	82.5	80.5	73.5	64.6
LIMO-32B	18.0	29.0	36.0	32.5	31.0	29.3	67.3
OpenThinker3-7B	81.5	87.0	89.0	84.5	86.0	85.6	72.1
R1-Distill-Qwen-32B	21.0	70.5	78.5	90.5	88.5	69.8	72.3
Qwen3-8B	71.0	88.5	89.0	91.5	89.5	85.9	79.1
QwQ-32B	53.0	79.5	86.5	88.5	91.0	79.7	80.5
Qwen3-32B	32.5	74.5	88.5	81.0	82.5	71.8	81.0
Qwen3-30B-A3B	68.0	90.5	93.5	91.5	95.5	87.8	81.1
AM-Thinking-32B	16.5	29.0	36.5	41.0	44.0	33.4	82.6

Table 5: **Recoverability (shared)** results (on 200 questions fully solved by all 15 LRMs eight out of eight). 0%, 20%, 40%, 60%, 80% are the positions of original reasoning where distraction is introduced. “Avg.” column averages across all the positions. “Benchmark Avg.” is from Table 4

Model	0%	20%	40%	60%	80%	Avg.	Benchmark Avg.
R1-Qwen-1.5B	89.0 $_{+45.0}$	94.0 $_{+28.0}$	91.0 $_{+27.0}$	89.5 $_{+22.5}$	84.0 $_{+22.0}$	89.5 $_{+28.9}$	47.5
R1-Llama-8B	95.5 $_{+30.0}$	96.5 $_{+15.0}$	97.0 $_{+12.5}$	91.5 $_{+9.0}$	87.0 $_{-6.0}$	93.5 $_{+12.1}$	54.1
DeepMath-1.5B	99.0 $_{+27.5}$	98.5 $_{+4.5}$	98.5 $_{+8.5}$	98.0 $_{+4.0}$	95.0 $_{+4.5}$	97.8 $_{+9.8}$	54.8
DeepScaleR-1.5B	97.0 $_{+35.5}$	97.5 $_{+9.5}$	97.5 $_{+8.0}$	98.0 $_{+13.0}$	86.0 $_{-2.0}$	95.2 $_{+12.8}$	53.3
OpenThinker3 1.5B	96.5 $_{+7.5}$	98.0 $_{+2.5}$	97.0 $_{+0.5}$	100.0 $_{+2.0}$	96.0 $_{-1.0}$	97.5 $_{+2.3}$	59.2
Qwen3-1.7B	100.0 $_{+3.0}$	100.0 $_{+0.5}$	100.0 $_{+1.0}$	100.0 $_{+1.5}$	82.0 $_{-16.0}$	96.4 $_{-2.0}$	59.9
R1-Qwen-7B	91.5 $_{+43.0}$	95.5 $_{+18.5}$	91.0 $_{+12.0}$	89.5 $_{+7.0}$	85.0 $_{+4.5}$	90.5 $_{+17.0}$	64.6
LIMO-32B	58.0 $_{+40.0}$	57.5 $_{+28.5}$	54.5 $_{+18.5}$	60.5 $_{+28.0}$	53.5 $_{+22.5}$	56.8 $_{+27.5}$	67.3
OpenThinker3-7B	93.0 $_{+11.5}$	94.5 $_{+7.5}$	96.0 $_{+7.0}$	96.5 $_{+12.0}$	85.0 $_{-1.0}$	93.0 $_{+7.4}$	72.1
R1-Qwen-32B	74.5 $_{+53.5}$	80.5 $_{+10.0}$	90.0 $_{+11.5}$	93.5 $_{+3.0}$	85.0 $_{-3.5}$	84.7 $_{+14.9}$	72.3
Qwen3-8B	95.5 $_{+24.5}$	97.0 $_{+8.5}$	97.5 $_{+8.5}$	97.0 $_{+5.5}$	80.0 $_{-9.5}$	93.4 $_{+7.5}$	79.1
QwQ-32B	64.5 $_{+11.5}$	73.0 $_{-6.5}$	81.0 $_{-5.5}$	90.0 $_{+1.5}$	86.5 $_{-4.5}$	79.0 $_{-0.7}$	80.5
Qwen3-32B	75.0 $_{+42.5}$	87.0 $_{+12.5}$	95.5 $_{+7.0}$	92.5 $_{+11.5}$	67.5 $_{-15.0}$	83.5 $_{+11.7}$	81.0
Qwen3-30B-A3B	83.5 $_{+15.5}$	88.0 $_{-2.5}$	91.0 $_{-2.5}$	94.0 $_{+2.5}$	66.0 $_{-29.5}$	84.5 $_{-3.3}$	81.1
AM-Thinking-32B	55.0 $_{+38.5}$	53.0 $_{+24.0}$	60.0 $_{+23.5}$	75.0 $_{+34.0}$	42.5 $_{-1.5}$	57.1 $_{+23.7}$	82.6

Table 6: Ablation Study: **Recoverability (shared)** results with original beginning (on 200 questions fully solved by all 15 LRMs eight out of eight). 0%, 20%, 40%, 60%, 80% are the positions of original reasoning where distraction is introduced. “Avg.” averages across all the positions. “Benchmark Avg.” is from Table 4

Table 7 and Table 8 report detailed breakdown of recoverability on individual subset; the former sets the length of distracting steer r^{steer} to be 0.2 times of the reasoning trajectory by default, whereas the latter sets to 0.4 of the reasoning trajectory.

Model	**0%**	**20%**	**40%**	**60%**	**80%**	**Avg.**	**Benchmark Avg.**
R1-Distill-Qwen-1.5B	24.0	40.8	40.8	38.8	48.4	38.6	47.5
R1-Llama-8B	32.0	38.4	49.2	57.6	79.8	49.6	54.1
DeepMath-1.5B	54.4	61.6	61.6	64.0	67.6	61.8	54.8
DeepScaleR-1.5B	35.2	54.0	56.8	57.6	60.8	52.9	53.3
OpenThinker3-1.5B	58.0	69.6	77.6	76.0	78.0	71.8	59.2
Qwen3-1.7B	58.4	70.4	74.4	85.2	84.4	74.6	59.9
R1-Distill-Qwen-7B	38.4	48.0	46.4	50.4	45.6	45.8	64.6
LIMO-32B	8.8	21.2	18.8	20.0	23.6	18.5	67.3
OpenThinker3-7B	63.2	72.4	76.4	77.6	82.8	74.5	72.1
R1-Distill-Qwen-32B	8.4	37.6	53.6	58.0	70.4	45.6	72.3
Qwen3-8B	51.6	64.4	73.2	76.0	78.8	68.8	79.1
QwQ-32B	50.0	54.5	64.8	68.8	74.8	62.6	80.5
Qwen3-32B	23.6	53.6	67.2	66.4	73.6	56.9	81.0
Qwen3-30B-A3B	36.8	61.6	68.8	67.6	65.2	60.0	81.1
AM-Thinking-32B	19.6	26.8	29.6	26.4	24.0	25.3	82.6

Table 7: **Recoverability-Random** results (on 200 randomly sampled questions for each of 15 LRMs). We sample questions according to the inverse proportions of solve rates. 0%, 20%, 40%, 60%, 80% are the positions of original reasoning where distraction is introduced. "Avg." averages across all the positions. "Benchmark Avg." is from Table 4

Model	**0%**	**20%**	**40%**	**60%**	**Avg.**	**Benchmark Avg.**
R1-Distill-Qwen-1.5B	11.6	26.0	27.6	24.0	22.3	47.5
R1-Llama-8B	29.2	43.2	54.8	56.4	45.9	54.1
DeepMath-1.5B	38.8	54.0	43.6	51.2	46.9	54.8
DeepScaleR-1.5B	24.8	50.0	53.2	50.4	44.6	53.3
OpenThinker3-1.5B	52.4	70.8	68.8	78.8	67.7	59.2
Qwen3-1.7B	59.2	73.2	76.4	81.2	72.5	59.9
R1-Distill-Qwen-7B	25.6	41.2	39.2	36.4	35.6	64.6
LIMO-32B	6.0	10.8	16.8	17.6	12.8	67.3
OpenThinker3-7B	59.6	72.0	70.0	73.2	68.7	72.1
R1-Distill-Qwen-32B	10.8	36.8	49.2	62.0	39.7	72.3
Qwen3-8B	50.4	67.2	71.2	76.0	66.2	79.1
QwQ-32B	44.8	52.0	61.2	68.4	56.6	80.5
Qwen3-32B	23.2	59.6	62.4	65.6	52.7	81.0
Qwen3-30B-A3B	31.6	53.2	62.0	59.6	51.6	81.1
AM-Thinking-32B	22.8	33.6	29.6	26.0	28.0	82.6

Table 8: **Recoverability-Random** results with **40% of distracting reasoning**. We control length of distraction to be 40% of distracting reasoning trace (default 20% in Table 5). The sampled questions are the same as in Table 5. 0%, 20%, 40%, 60% are the positions of original reasoning where distraction is injected. "Avg." averages across all positions. "Benchmark Avg." is from Table 4

E Guidability Test

Table 9 reports the number of unique problems and guiding trajectories used per guiding model (sub-column) for each LRM (row). Table 10 reports guidability (individual) results for different length of the guiding steers measured by $x\%$ of the trajectories. Similarly, Table 11 reports breakdown of guidability on shared subset. Table 12 groups guidability (individual) scores by the guiding models (column) for each LRM (row)

	# of Problems			# of Trajectories		
	DeepSeek-R1	**Qwen-3**	**QwQ-32B**	**DeepSeek-R1**	**Qwen-3**	**QwQ-32B**
DeepMath-1.5B	152	198	302	231	268	302
DeepScaleR-1.5B	154	196	311	234	269	311
LIMO-Qwen-32B	100	137	185	142	172	185
OpenThinker3-1.5B	151	199	270	236	278	270
OpenThinker3-7B	101	146	163	146	186	163
Qwen3-1.7B	130	175	245	192	233	245
R1-Distill-Llama-8B	151	196	266	229	269	266
R1-Distill-Qwen-1.5B	168	213	363	261	290	363
R1-Distill-Qwen-7B	107	156	190	151	195	190
R1-Distill-Qwen-32B	94	145	162	134	182	162

Table 9: **Guidability statistics**: unique number of problems and trajectories per guiding model (column) for different student models (row) for **Guidability (individual)** test.

Model	**20%**	**40%**	**60%**	**80%**	**Avg**	**Benchmark Avg.**
R1-Distill-Qwen-1.5B	$14.6_{7.7}$	$23.1_{17.2}$	$33.2_{31.3}$	$43.0_{46.2}$	$28.4_{25.6}$	47.5
R1-Distill-Llama-8B	$20.8_{5.4}$	$29.6_{15.7}$	$40.0_{27.6}$	$49.7_{34.8}$	$35.0_{21.8}$	54.1
DeepMath-1.5B	$13.6_{7.2}$	$21.1_{16.2}$	$31.2_{27.5}$	$42.3_{40.6}$	$27.1_{22.9}$	54.8
DeepScaleR-1.5B	$15.7_{7.5}$	$23.2_{15.7}$	$34.6_{28.1}$	$45.6_{41.8}$	$29.8_{23.3}$	53.3
OpenThinker3-1.5B	$18.1_{11.0}$	$30.6_{21.4}$	$36.1_{32.3}$	$46.0_{42.3}$	$32.7_{26.9}$	59.2
Qwen3-1.7B	$18.2_{5.8}$	$23.7_{11.8}$	$34.8_{20.6}$	$42.8_{33.8}$	$29.9_{18.0}$	59.9
R1-Distill-Qwen-7B	$10.8_{3.5}$	$16.2_{6.3}$	$22.0_{13.1}$	$29.9_{25.4}$	$19.7_{12.1}$	64.6
LIMO-32B	$12.6_{2.6}$	$18.8_{4.8}$	$24.4_{11.6}$	$30.0_{21.8}$	$21.5_{10.2}$	67.3
OpenThinker3-7B	$11.1_{6.5}$	$20.0_{10.1}$	$22.6_{15.4}$	$28.7_{23.4}$	$20.6_{13.8}$	72.1
R1-Distill-Qwen-32B	$14.2_{3.8}$	$19.7_{6.1}$	$24.9_{12.4}$	$31.2_{22.6}$	$22.5_{11.2}$	72.3

Table 10: **Guidability (individual)** results (on all questions with solve rate $\leq \frac{1}{8}$ **for each individual model**). 20%, 40%, 60%, 80% are proportion of teacher reasoning revealed to the student model in its thinking window. The subscript value is the percentage of cases where teachers **have derived the solution**. "Avg" is the average across different proportions. "Benchmark Avg" is the same as in Table 4.

Model	20%	40%	60%	80%	Avg	Benchmark Avg.
R1-Distill-Qwen-1.5B	1.2	0.9	4.1	5.8	3.0	47.5
R1-Distill-Llama-8B	5.2	5.8	10.4	13.3	8.7	54.1
DeepMath-1.5B	0.9	0.9	4.6	7.2	3.4	54.8
DeepScaleR-1.5B	1.2	0.9	5.2	9.0	4.1	53.3
OpenThinker3-1.5B	1.7	5.5	7.0	8.4	5.7	59.2
Qwen3-1.7B	2.3	3.2	7.8	11.0	6.1	59.9
R1-Distill-Qwen-7B	2.6	5.2	6.4	9.9	6.0	64.6
LIMO-32B	4.9	7.5	10.1	12.8	8.8	67.3
OpenThinker3-7B	4.9	9.0	9.6	12.8	9.1	72.1
R1-Distill-Qwen-32B	4.1	7.5	11.0	14.2	9.2	72.3

Table 11: **Guidability (shared)** results (on questions with solve rate $\leq \frac{1}{8}$ **across all ten models**). 20%, 40%, 60%, 80% are proportion of teacher reasoning revealed to the student model in its thinking window. "Avg" is the average across different proportions. "Benchmark Avg" is the same as in Table 4.

Model	DeepSeek-R1	QwQ-32B	Qwen3-235B-A22B	Benchmark Avg.
R1-Distill-Qwen-1.5B	28.2	30.4	26.2	47.5
DeepMath-1.5B	29.0	26.2	26.3	54.8
DeepScaleR-1.5B	30.9	31.1	27.3	53.3
R1-Distill-Llama-8B	37.8	34.4	33.2	54.1
Qwen3-1.7B	33.4	31.1	25.6	59.9
OpenThinker3-1.5B	35.7	30.6	32.3	59.2
R1-Distill-Qwen-7B	22.0	19.6	18.7	64.6
LIMO-32B	24.5	24.6	15.7	67.3
R1-Distill-Qwen-32B	23.5	23.0	21.9	72.3
OpenThinker3-7B	22.9	21.4	18.0	77.8

Table 12: **Guidability (individual)** results (teacher model comparison). Each teacher model averages across **Guidability (individual)** scores for all proportions, 20%, 40%, 60%, 80%, in Table 10

F Control Study

Supervised Fine-Tuning Hyperparameters. We perform full fine-tuning on `Qwen2.5-1.5B` and `Qwen2.5-3B` base models for 5 epochs. The max tokens is set to 16K, batch size 64, learning rate 2e-5, warmup ratio 0.1, max gradient norm 1.0, weight decay 0.01.

Ablation Study. We compare the effects of distillation teachers on `Qwen2.5-7B` models. We observe similar patterns as discussed in §4.1, where AM-Distill models achieve worse recoverability compared to QwQ-/Qwen3-Distill models. The guidability scores are not measured since the benchmark performance are too high to collect sufficient qualified problems.

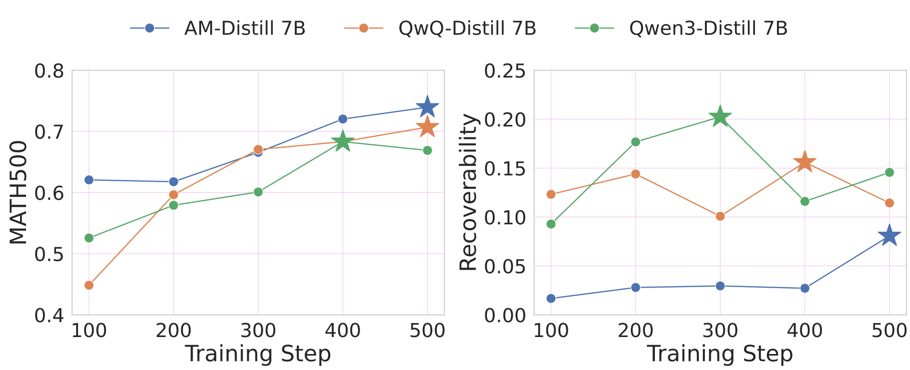

Figure 8: Qwen2.5 7B models distilled from AM (Thinking-v1) 32B also shows lower recoverability than those distilled from QwQ 32B or Qwen 32B, while having similar benchmark performance; the gap is **significant for all steps** (p $\leq$ 0.005). Stars mark each model's peak over training steps.