Measuring Off-Trajectory Math Reasoning of LLMs

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

Reasoning LLMs are trained to verbalize their thinking process, yielding strong gains on math benchmarks. This transparency also opens a promising direction: multiple reasoners can directly collaborate on each other's thinking on a shared trajectory, yielding better inference efficiency and exploration. A key prerequisite, however, is the ability to assess usefulness 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 yield desired *off-trajectory* behaviors? We propose twin tests that capture the two extremes of the off-trajectory spectrum, namely **Recoverability**, which tests whether LLMs 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 LLMs (1.5B–32B) and reveals a counterintuitive finding—"stronger" LLMs 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%. This work lays the groundwork for evaluating multi-model collaborations under shared reasoning, while revealing limitations of off-the-shelf reasoning LLMs.

1 Introduction

LLMs with thinking abilities (e.g., OpenAI's o-series [21], DeepSeek-R1 [13]) have recently emerged as the frontier models for math reasoning tasks. These models, trained with reinforcement learning with verifiable rewards (RLVR) [38] or distillation [18], learn to verbalize their intermediate reasoning in language [11]. This paradigm has promising implications, e.g., better **efficiency** (large LLMs can delegate easy derivations or arithmetic checking to smaller models) and broader **exploration** (models with complementary expertise can expand the reasoning search by spawning and combining diverse branches) [1, 5, 7, 34, 33].

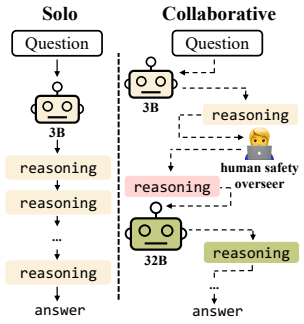

Figure 1: Comparison of solo (left) vs. collaborative reasoning (right).

Most LLMs today are trained and evaluated to reason 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? Ideally, LLMs should integrate useful insights and backtrack from erroneous steps made by collaborators, even when these traces are not in-distribution. We call this ability *off-trajectory reasoning* and ask: **can solo-reasoning LLMs collaborate conditioned on out-of-distribution trajectories?**

We decompose off-trajectory reasoning into two parts, **recoverability** and **guidability**, and evaluate both in simulated collaboration scenarios (see Figure 2). The recoverability test evaluates how well

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

LLMs can reject a collaborator's irrelevant thinking or erroneous steps and return to the original correct path. At the other end of the spectrum, the guidability test evaluates whether LLMs continue from another model's correct partial derivation to solve problems that solo-reasoning fails.

We systematically evaluate 15 open-weight LLMs on a suite of five math benchmarks [30, 31, 17, 27, 14]. Surprisingly, we find that stronger reasoning models are more prone to failure under off-trajectory distractions. Under the recoverability test, their performance falls by **25.1%**. At the same time, guidability test reveals that LLMs fail to continue from other models' correct trajectories, even when correct answers appear in these trajectories. Overall, our results show that LLMs can neither reject distracting nor build upon useful off-trajectory inputs. Moreover, the practice of benchmark over-optimization fails to capture broader reasoning capabilities, of which collaborative and off-trajectory reasoning is an integral part.

2 Twin Tests for Off-Trajectory Math Reasoning

Preliminaries and Notations. Let M be a reasoning model and (q, a^*) be a datapoint. In solo-reasoning, M generates a trajectory $\mathbf{r} = [r_1, \ldots, r_k]$ and answer a for a math question q, i.e. $(\mathbf{r}, a) \sim M(\cdot \mid q)$. r_i refers to a *reasoning unit* of any granularity (step, sentence, etc.). In the collaborative setting, multiple models contribute to parts of the trajectory $\mathbf{r}$. The main model M needs to build upon a trajectory mixing both in- and out-of-distribution reasoning units $\mathbf{r} = [r^M, ..r^{M'}, r^{M''}, \ldots, r^M]$. In this paper, we study a simplified setup of two-model collaboration for math reasoning.

Two-model Setup We simulate collaboration between the main model M and a collaborator M_{steer}, which together form an off-trajectory reasoning $[r^{\text{og}}, r^{\text{steer}}]$. In practice, we sample r^{og} from M for the first m tokens and r^{steer} from M_{steer} for the first n tokens. Then, we prompt M to complete conditioned on the question and the steered trajectory, i.e., $(\mathbf{r}^{\text{off}}, a^{\text{off}}) \sim M(. \mid q, [r^{\text{og}}, r^{\text{steer}}])$. We can measure the success based on answer correctness i.e. $\mathbb{E}_{\{(q,a^*)\sim\mathcal{D}\}}[\mathbb{1}[a^{\text{off}} = a^*]]$.

Considerations for designing the steer. Our setup lets us simulate two extremes of how a steer affects the main model M. A steer can be *distracting*, misleading M away from its original correct trajectory, or *guiding*, offering partial reasoning that helps M solve problems beyond its solo ability. Based on these desiderata, we design twin tests—(i) **Recoverability**: can M resist a distractor and backtrack to its original correct trajectory? (ii) **Guidability**: can M effectively leverage a guiding steer from a stronger reasoning model to surpass its solo-reasoning ability? These twin tests differ in (i) the selection of test questions q and (ii) the construction of steered trajectories $[r^{\text{og}}, r^{\text{steer}}]$. Given a dataset $\mathcal{D}$ and test model M, our protocol automatically instantiates M-specific off-trajectory dataset for both tests, i.e. $\mathcal{D}_M^{\text{test}} = \{(q, [r^{\text{og}}, r^{\text{steer}}], a^*)\}$. (See Figure 2)

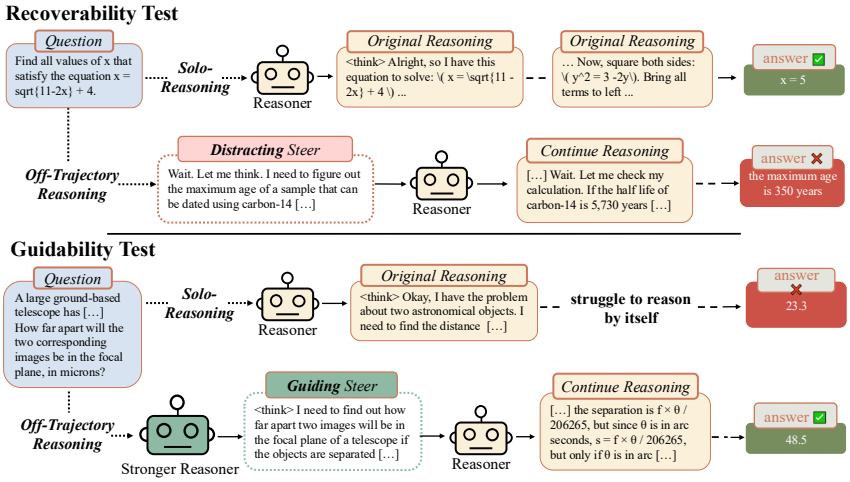

Figure 2: Illustration of the twin tests: we perturb a model's reasoning trajectories with off-trajectory steers to evaluate its *recoverability* (under a distracting steer) or *guidability* (under a guiding steer). The distracting steer is sampled from the same reasoner but for a different question.

2.1 Recoverability Test

Selecting test datapoints $\{(q, a^*)\}$. For a given test model M, we select the subset of test questions which M can correctly answer in solo-reasoning, i.e. $a = a^*$, where $(r, a) \sim M(. \mid q)$. 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 want r^{steer} to be a strong distractor for M. However, it is hard to determine *a priori* which collaborator M_{steer} and steer r^{steer} will reliably do so. To ensure distraction, we instead sample r^{steer} from M itself, but conditioned on a different question q'. So, if M blindly completes from r^{steer}, its reasoning is guaranteed to be wrong. 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. $|r^{og}|$.

2.2 Guidability Test

Selecting test datapoints $\{(q, a^*)\}$. The goal is to test whether M can leverage a *guiding steer*, i.e., a correct partial reasoning, on questions it fails in solo reasoning. We therefore select questions where the solo-reasoning solve rate of M is 0 or 1 out of 8 samples, at its capability boundary.

Constructing steered trajectories. Unlike the recoverability test, we set $m = 0$ and exclude M's own partial reasoning, since r^{og} might already contain errors that anchor M in a wrong direction, which could confound the measurement of guidability.

The steer r^{steer} is drawn from a stronger reasoner M_{steer} with higher benchmark performance than M. To test how well M can build on the guiding steer r^{steer}, we truncate the sampled trajectory to the first n tokens. We vary n to control the "amount" of guidance provided to M, and we use multiple guiding models M_{steer} to generate independent steers for each question q, which enables guidability measurement under different steer distributions and amount of guidance.

3 Off-the-shelf Evaluation & Results

3.1 Experiment Setup

LLMs and Math Benchmarks. We run our experiments on 15 open-weight models, grouped into four families—(1) **DeepSeek-R1** [13]: `R1-Qwen-1.5B/7B/32B` and `R1-Llama-8B`; (2) **Qwen3** [42]: `Qwen3-1.7B/8B/30B-A3B` and `Qwen3-32B`; (3) **QwQ**: `QwQ-32B` and `OpenThinker3-1.5B/7B` [12]; (4) **Community**: `DeepScaleR-1.5B` [29], `DeepMath-1.5B` [16], `LIMO-32B` [43], and `AM-Thinking-32B` [23]. We include more detailed model information in Appendix A. We evaluate on a pool of 1,507 math questions sourced from five standard benchmarks, AIME-2024 [30], AIME-2025 [31], MATH-500 [17], Minerva (math subset) [27], and OlympiadBench [14].

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

Recoverability and Guidability Setup. For **recoverability test**, we sample 200 original trajectories r^{og} and 50 distracting trajectories r^{steer} per LLM. By default, we set n, i.e. $|r^{\text{steer}}|$ to be 0.2 times the trajectory length, leaving enough tokens for off-trajectory completion. We vary the length of r^{og} as $\{0, 0.2, 0.4, 0.6, 0.8\}$ of the reasoning from the main model. Recoverability is reported on two subsets: (1) *shared*—questions all 15 LLMs solve perfectly (8 out of 8), and (2) *individual*—questions selected per model as defined in §2.

For the **guidability test**, we select `DeepSeek-R1`, `Qwen3-235B`, and `QwQ-32B` as M_{steer}. Since the top 5 LLMs almost saturate the benchmarks, we evaluate only the other 10 with enough questions with solve rate $\leq$ 1/8 (Table 2). Steer length n, i.e., $|r^{\text{steer}}|$, is set to $\{0.2, 0.4, 0.6, 0.8\}$ of the trajectory. Guidability is also reported on *shared* (intersection across models) and *individual* (per model) subsets.

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.

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$_{\uparrow5}$	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 LLMs from four families.** Columns report benchmark averages and recoverability/guidability scores for *shared* (Sh.) and *individual* (Ind.) subsets. 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. N/A = not evaluated.

Finding 1: Stronger solo-reasoners $\neq$ stronger collaborators. Surprisingly, recoverability and guidability are largely orthogonal to solo-reasoning. In particular, models in the *low* benchmark tier (e.g., `OpenThinker3-1.5B` and `Qwen3-1.7B`) show substantially better recoverability than *medium* and *high* tier models like `QwQ-32B` and `Qwen3-32B`. The best solo-reasoning model `AM-Thinking-32B` reports the second-worst recoverability. 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 a mean 25.1% degradation in reasoning performance in the recoverability test. In addition, all LLMs report exceptionally low guidability scores; none of the evaluated models report $>10\%$ on the shared subset. Taken together, these findings suggest that **models optimized heavily for popular math benchmarks may have hidden vulnerabilities, particularly in off-trajectory reasoning**.

Finding 2: The beginning paragraph of reasoning is critical for recovery. Table 1 visualizes the recovery rates when distracting steers are inserted at different positions (%) of the original trajectory. Across models, distraction at the very start (0%) causes the largest degradation. This is surprising since the opening of reasoning usually only restates the question and rarely includes actual problem solving. So, we hypothesize that restating the question at the start is critical for anchoring later reasoning.

To test this, we conduct an ablation that re-instantiates the recoverability while preserving the first paragraph of the original trajectory. Most LLMs experience noticeable improvements, especially at 0% (See Table 4). With this small tweak, average recoverability exceeds 83.5% for all models except `LIMO-32B` and `AM-Thinking-32B`). This shows that **while restatement of the problem does not add new information, it is critical for LLM off-trajectory reasoning**.

Finding 3: LLMs fail to leverage guidance to surpass their inherent limits. Table 1 shows that all models, regardless of their solo-reasoning abilities, struggle to build upon guiding trajectories. To our surprise, their guidability does not improve even when models are paired with their own distillation teacher (see Table 5 for full set of results). For example, `Qwen3-1.7B` shows no guidability gains when guided by `Qwen3-235B` compared to other models.

Furthermore, we find that **already low guidability scores are partly inflated**. Since the guiding steers are truncated at different lengths, on average 18.6% of them already contain the full correct derivation and answer (See *Ans.?* column Table 7 and breakdown in Table 6). In such cases, we expect the guidability test to be trivially easy, yet we find that LLMs 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 models on correct but out-of-distribution traces does not enable them to successfully leverage them and surpass their inherent capability limits.

4 Related Work

Large Reasoning Models. Recent post-training advances have led to massive improvements on math and coding benchmarks [20, 13], as exhibited by both closed- and open-source LLMs since the release of OpenAI's o-1 [21], e.g., [13, 42, 12, 43, 23]. These models are trained to produce extended reasoning traces using RL algorithms such as Proximal Policy Optimization (PPO) [37], Grouped Relative Policy Optimization (GRPO), and related variants [38], typically with verifiable rewards. At smaller scales (under 32B parameters), reasoning models like R1-Qwen-Distill series [13] and Qwen3 family [42] are primarily trained with distillation [18]. 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 LLMs spanning diverse model families and training paradigms.

LLM Reasoning Intervention and Collaboration. Recent studies intervene on the LLM reasoning process to understand and control their behaviors, including perturbing intermediate steps to examine their faithfulness [2, 3], improve instruction following and alignment behaviors [41], or interpret [26, 32] and stress-test cognitive behaviors [11]. [40] examines the impact of thinking patterns on outcome correctness, while [15, 26] systematically categorize different types of reasoning strategies and errors. In addition, our work fits within prior work on teacher–student framework for augmenting model reasoning [19, 1, 4]. In a closely related work, [15] investigates LLMs' 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 [33], including offloading challenging reasoning parts to larger models [1] and orchestrating different models for high-level planning and downstream execution [25]. Our work evaluates how solo-reasoning LLMs can fail when routed onto a shared reasoning trajectory.

5 Limitations & Future Work

Our study conducts an initial systematic investigation into the fragility of LLM 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 LLMs 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 [22, 24, 6], science [39, 36, 10], and logical reasoning tasks [9, 28, 8].

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 LLMs relative to distractors from other models.

6 Conclusion

In this work, we study off-trajectory reasoning in LLMs—their ability to recover from or build on reasoning steered by other models. We propose twin tests: Recoverability, which measures whether models can backtrack from distracting steers, and Guidability, which measures how well they can take advantage of guiding steers. Across 15 open-weight LLMs, our evaluation reveals consistently poor performance on both, underscoring the limitations of solo-reasoning LLMs in collaborative settings and pointing to directions for future work to advance multi-model math reasoning systems.

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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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/`.

[15] 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.

[16] 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.

[17] 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`.

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

[19] 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/.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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`.

A LLM Selection Details

Below are the list of LLMs we evaluate in this study with their detailed information. Table 2 reports the number of unique problems and guiding trajectories used per guiding model (sub-column) for each LLM (row).

- **DeepSeek-R1** [13]: `R1-Qwen-1.5B/7B/32B` and `R1-Llama-8B` are distilled from `DeepSeek-R1` using supervised fine-tuning (SFT).
- **Qwen3** [42]: `Qwen3-32B` is trained using RL for reasoning without distillation, while `Qwen3-1.7B/8B/30B-A3B` are distilled from `Qwen3-235B` and `Qwen3-32B`.
- **QwQ**: `QwQ-32B` [35] is trained from `Qwen2.5-32B-Base` model with RL. `OpenThinker3-1.5B/7B` [12] are based on `Qwen2.5-Instruct` and distilled from `QwQ-32B` on 1.2M curated math and coding examples.
- **Community**: `DeepScaleR-1.5B` [29] and `DeepMath-1.5B` [16] are trained using RL on `R1-Qwen-1.5B` using DeepScaleR and DeepMath datasets respectively. `LIMO-32B` [43] is SFT from `Qwen2.5-32B-Instruct` on the LIMO dataset of 817 examples. Finally, `AM-Thinking-32B` [23] is a `Qwen2.5-32B-Base` model first distilled on 2.84M examples and RL on 54K math and coding questions.

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

B Details for Finding 2 & 3

Table 4 reports the results of ablation study explained in §3.2, where the first paragraph of model reasoning is preserved. The subscripts in Table 4 equals the difference between the major numbers in Table minus the corresponding numbers in Table 3 to show the changes in recoverability induced by the small tweak in trajectory. Table 5 groups guidability (individual) scores by the guiding models (column) for each LLM (row). Table 6 reports guidability (individual) results for different length of the guiding steers measured by $x\%$ of the trajectories.

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 3: **Recoverability (shared)** results (on 200 questions fully solved by all 15 LLMs 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.

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 4: Ablation Study: **Recoverability (shared)** results with original beginning (on 200 questions fully solved by all 15 LLMs 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.

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 5: **Guidability (individual)** results (teacher model comparison). Each teacher model averages across **Guidability (individual)** scores for all proportions, 20%, 40%, 60%, 80%, in Table 6

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 6: **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.

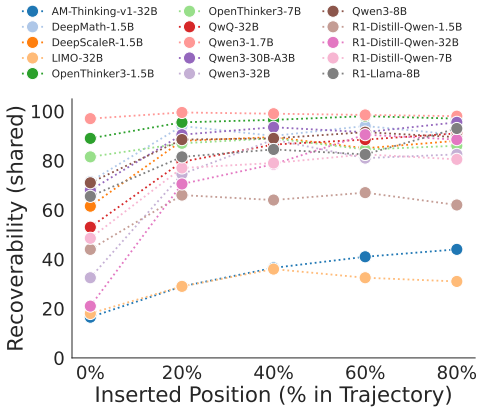

Figure 3: Recoverability (shared) across positions (%) of the original trajectory for 15 LLMs

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