A Primer in Post-Training Reasoning Data: What We Know About How It Works

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training has become a major driver of recent progress in large reasoning models, and reasoning data are often a key determinant of its success. Research on these data has grown rapidly, yet the evidence remains fragmented. This primer synthesizes over 150 works around four questions: what forms reasoning data take, what makes them useful, how they are constructed, and what scaling studies establish. Across mathematics, code, science, and interactive tasks, we compare apparently conflicting findings to clarify the conditions under which conclusions hold and the questions that remain open. We distill this evidence into seven practical lessons and reporting recommendations to guide data construction, assess feedback reliability, and evaluate claims of data efficiency and capability improvement. We conclude by outlining directions for future research.

1 INTRODUCTION

The training of large language models usually has two stages: pre-training and post-training. With the rise of o1-style test-time scaling and reasoning models, post-training matters more and more for advancing model capabilities (OpenAI, 2024; DeepSeek-AI et al., 2025). Recent studies show that data quality and construction can greatly affect performance even when the base model and training recipe are held fixed (Muennighoff et al., 2025; Guha et al., 2025). Yet findings about how post-training reasoning data are built, curated, and used remain scattered across studies. This primer synthesizes over 150 works to connect these findings and sum up their practical implications.

However, post-training reasoning data are not just collections of datasets, but evidence-bearing records for explaining how models improve. As Table 1 summarizes, several intuitive assumptions about reasoning data can be misleading: correct final answers do not guarantee correct reasoning, with step errors in 51.8% of correct-answer Omni-MATH solutions in ProcessBench (Zheng et al., 2025); harder problems may be filtered out by RL when all sampled answers fail, yet still help on-policy distillation (Yu et al., 2025; Hou et al., 2026); larger datasets can add repeated answers without widening task coverage (Guha et al., 2025; He et al., 2026); stronger solvers do not always make better teachers (Guha et al., 2025); and overly cleaned success trajectories can drop failures that serve as negative examples or context for learning recovery (Pan et al., 2025; Xu et al., 2025b).

A recurring principle emerges: correctness alone does not determine teaching value; the learner and how supervision is used also matter. Our contributions are threefold. (i) We organize the evidence around four questions—what forms reasoning data take, what makes them useful, how they are built, and what scaling studies establish (Sections 2–6)—and distill seven practical lessons with reporting recommendations (Table 1). (ii) We trace six of seven apparent disagreements to differences in the learner or the use of supervision, distinguishing plausible explanations from controlled evidence (Appendix B). (iii) We propose three controlled comparisons that test predictions on teacher selection, problem difficulty, and feedback reliability (Section 7).

2 WHAT IS POST-TRAINING REASONING DATA?

What a record contains. Most public reasoning datasets pair a question with a reasoning trace and a final answer, from the 1,000 examples of s1K (Muennighoff et al., 2025) to the 1.2 million of

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

Counterintuitive lessons about reasoning data	
🪃 **Trap: "A correct answer means correct reasoning."** (Zheng et al., 2025; Turpin et al., 2023)	
💡 **Takeaway**	Experts find step errors in 51.8% of correct-answer Omni-MATH solutions, and traces can omit what drove the answer (Sec. 4.2).
🚀 **Future Reporting**	Report how the steps were checked, not only the final answer.
🪃 **Trap: "Harder data is better data."** (Yu et al., 2025; He et al., 2025a; Hou et al., 2026)	
💡 **Takeaway**	Difficulty depends on the learner: problems it never solves give RL no signal, yet often teach most in on-policy distillation (Sec. 4.3).
🚀 **Future Reporting**	Report the learner, samples per problem, pass-rate band and update rule.
🪃 **Trap: "More data means better coverage."** (Guha et al., 2025; He et al., 2026; Shumailov et al., 2024)	
💡 **Takeaway**	A row count mixes new questions, new answers and repeats; new task types widen coverage, while rewriting a pool mostly reweights it (Sec. 4.4).
🚀 **Future Reporting**	Report distinct questions, answers per question and task types.
🪃 **Trap: "The strongest solver is the best teacher."** (Muennighoff et al., 2025; Guha et al., 2025)	
💡 **Takeaway**	QwQ-32B trails DeepSeek-R1 on benchmarks yet trains a stronger student; good traces are ones this student can learn from (Sec. 5.2).
🚀 **Future Reporting**	Report the trace writer, the student and the question set together.
🪃 **Trap: "A clean successful transcript is ideal."** (Pan et al., 2025; Xu et al., 2025b; Chen et al., 2025c)	
💡 **Takeaway**	Failed attempts teach as negatives, as partial credit or as masked context, as long as they are not imitated (Sec. 5.3).
🚀 **Future Reporting**	Release failed attempts and state how each one enters the loss.
🪃 **Trap: "Self-play removes curation."** (Zhao et al., 2025a; Huang et al., 2026; Pan et al., 2026)	
💡 **Takeaway**	Curation moves into the check that decides correctness; under majority vote, R-Zero's label accuracy falls from 79% to 63% (Sec. 5.4).
🚀 **Future Reporting**	Report what decides correctness and the label accuracy at each iteration.
🪃 **Trap: "A higher score means a better model."** (Hochlehnert et al., 2025; Zuo et al., 2025)	
💡 **Takeaway**	Scores shift with seeds, decoding, extra thinking and training on the test questions themselves; a gain need not mean better data (Sec. 6.3).
🚀 **Future Reporting**	Report seeds, decoding settings, inference budget and test-set exposure.

Table 1: **Counterintuitive lessons about reasoning data.** Each trap is a common intuition; the takeaway gives the evidence and its section, and the last row what reports should state.

OpenThoughts3, which pairs about 75K questions with 16 traces each (Guha et al., 2025); we call such a unit a *record*. A record may also carry *feedback*, such as the step-level labels of PRM800K (Lightman et al., 2023), and *provenance*, meaning who produced each part, which is often poorly documented (Longpre et al., 2024). Figure 1a shows one record for the Game of 24 puzzle "3, 3, 8, 8": a three-step trace, the answer $8 \div (3 - 8/3) = 24$, a checker's verdict, and a note on who produced it.

What makes it reasoning data. The literature separates reasoning data from general instruction data (Ouyang et al., 2022) by two properties. The first is the trace. Wei et al. (2023) define a chain of thought as a series of intermediate reasoning steps, and Chen et al. (2025a) characterize the longer chains of o1-style models by deep reasoning, extensive exploration and reflection. Several studies find that such traces teach through structure more than content. On the Countdown arithmetic game, Gandhi et al. (2025) show that solutions displaying verification and backtracking prepare a model for RL even when their answers are wrong. Li et al. (2025) find that training on traces with wrong final answers lowers average accuracy by only 3.2 points, whereas shuffling the steps costs 10.2 points and leaves the model below where it started. The second property is a checkable endpoint. Lambert et al. (2025) name training on such checks reinforcement learning with verifiable rewards (RLVR), and DeepSeek-R1 scales it with simple rule-based rewards (DeepSeek-AI et al., 2025). A checkable endpoint need not be a reference string: in Figure 1a, a program confirms that the answer equals 24 and uses each number once. RL datasets therefore select for checkability: Big-Math keeps only problems with a unique, closed-form answer (Albalak et al., 2025), and DAPO rewrites problems so that every answer is an integer (Yu et al., 2025).

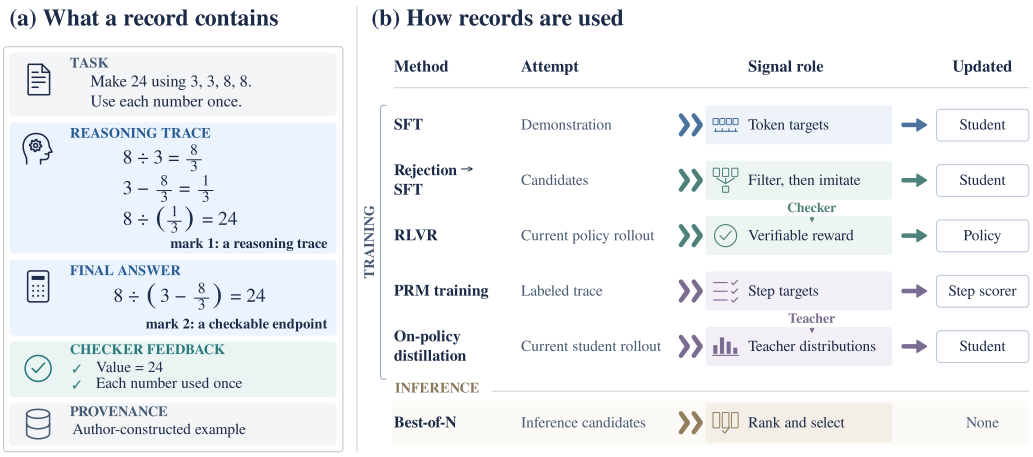

Figure 1: **One record, several learning interfaces.** (a) An illustrative Game of 24 record, a special case of Countdown (Gandhi et al., 2025); its trace and checkable endpoint mark it as reasoning data. (b) Common recipes draw different signals from it: blue from the trace and answer, green from the checker, purple from supervision the record lacks. Only best-of-N updates no model.

One record, several uses. The same record can enter training in several ways (Figure 1b). In supervised fine-tuning (SFT), its trace and answer are imitated directly (Ouyang et al., 2022), as in s1K and OpenThoughts3. A checker can first filter a model's own attempts, which are then imitated, as in STaR (Zelikman et al., 2022), ReST^{EM} (Singh et al., 2024) and the second stage of DeepSeek-R1 (DeepSeek-AI et al., 2025); the same checker can instead serve as the reward in RLVR, as in DAPO (Yu et al., 2025). Other uses need supervision the record lacks: step labels train a process reward model (PRM) (Lightman et al., 2023; Wang et al., 2024), and a teacher's next-token distributions on the student's own attempts drive on-policy distillation (Agarwal et al., 2024). A scorer can also rank candidates at inference time without updating any model (Cobbe et al., 2021; Lightman et al., 2023). Releases often anticipate several uses at once: DeepMath-103K ships each problem with a verifiable answer for RL and three DeepSeek-R1 solutions for SFT (He et al., 2025b). We call the model that is updated the *learner*: the student in distillation, the policy in RL. Training is *on-policy* when the learner made the attempt and *off-policy* otherwise; we call the combination of learner, target and update rule a record's *learning interface*.

Scope and evidence. We cover mathematics, code, formal proof and tool use, plus efforts that extend reasoning data to other domains with reference answers (Yuan et al., 2025a). The evidence comes from over 150 studies, datasets and system reports (Appendix A.3) and is densest in mathematics. Because reasoning benchmarks are sensitive to seeds and decoding settings (Hochlehnert et al., 2025), we flag comparisons that change more than one thing at once.

3 What Forms Does Reasoning Data Take?

What a record retains. Records range from questions paired with reference answers, as in Big-Math (Albalak et al., 2025), to reasoning traces and full interaction histories. DeepMath-103K pairs each problem with a verifiable answer and three R1 solutions (He et al., 2025b), while SWE-Gym provides trajectories of agents working on repository tasks (Pan et al., 2025). Records can additionally carry step-level feedback: Math-Shepherd labels solution steps automatically using continuations sampled from each step to assess whether the correct answer remains reachable (Wang et al., 2024). Such labels add supervision to a trace rather than define a separate level of record completeness.

How feedback is produced. Three common approaches produce feedback, with overlapping implementations and different limitations. *Programmatic* checks compare answers, run unit tests, or verify formal proofs; KodCode pairs coding questions with solutions and unit tests (Xu et al., 2025d). Final-answer checks alone do not establish the validity of intermediate reasoning: Yuan et al. (2025b) find solutions that reach the correct answer through unsound steps. *Environmental*

checks evaluate the state an interaction leaves behind, often through programmatic tests. AppWorld checks task completion and specified constraints on unintended state changes (Trivedi et al., 2024). Agents can nonetheless pass the visible tests while missing the intended goal (Zhao et al., 2026). *Judgment-based* checks assess responses against rubrics or preferences, and rubric scores can serve directly as RL rewards (Gunjal et al., 2025). LLM judges can, however, exhibit position, verbosity, and self-preference biases (Zheng et al., 2023).

Combining forms and checks. Record form does not prescribe a single feedback procedure or learning use. SWE-Gym's test-checked trajectories train both an agent and a verifier that ranks candidate solutions (Pan et al., 2025). Multiple checks can also be applied to the same attempt: Jain et al. (2025) find that test-based and model-based verifiers fail in complementary ways, and that combining them outperforms either alone in their setting.

Takeaway. Reasoning data vary along two dimensions: what a record retains and how feedback is produced. A full reasoning trace may receive only an answer-level check, while the same record can be assessed using multiple checks with different coverage.

4 What Makes Reasoning Data Useful?

Good reasoning data are usually described by four properties: correct labels, sound traces, suitable difficulty and broad coverage. The evidence supports each one, but always relative to something else: a checker, a learner or a budget (Figure 2).

4.1 Checkers Make Mistakes in Both Directions

Checker errors change what a model learns, in both directions. Huang et al. (2025) find that rule-based math verifiers often reject correct answers in unexpected formats, and that these false negatives hurt RL more as the policy improves. Xu et al. (2025c) estimate that over 38% of Big-Math responses rejected by a rule-based verifier are correct; recovering them raises accuracy by up to 10 points on a hard-to-verify benchmark. Model-based verifiers err the other way: Zhao et al. (2025b) show that open-source LLM judges accept responses that are only a colon or "Thought process:" up to 94% of the time, and a colon fools even Claude-4-Sonnet in over half the cases on MATH. Even proof assistants rely on error-prone formalization: only 60% of automatic formalizations initially run in Lean, rising to 87% after refinement (Leang et al., 2025).

There is conflicting evidence, however, about how much label accuracy matters (Figure 2a). Shao et al. (2026) report that even random rewards lift Qwen2.5-Math-7B by 21.4 points on MATH-500, against 29.1 for correct rewards, but hurt Llama3.1-8B-Instruct and OLMo2-7B. Yan et al. (2026) suggest that such gains retrieve memorized solutions rather than build new reasoning, consistent with signs that Qwen2.5-Math has seen MATH-500 problems during pre-training (Wu et al., 2025). **A checker's verdict and a label's usefulness are therefore different things**: label accuracy seems to matter most when training must teach the model something it does not already know.

4.2 Right Answers Can Hide Wrong Steps

A correct final answer does not certify the steps behind it. Among correct-answer solutions, Process-Bench's expert annotators find step errors in 3.5% of GSM8K solutions and 51.8% of Omni-MATH solutions (Figure 2b; Zheng et al., 2025), and PRMs that accept such solutions inflate best-of-N scores (Zhang et al., 2025b). A solution to $2x + 3 = 7$ that writes $2x = 5$ but continues from $2x = 4$ still ends at $x = 2$. Even locating the first wrong step is hard: on the same benchmark, the best open PRM reaches an average F1 of 56.5, against 71.5 for QwQ-32B-Preview prompted to critique each step (Zheng et al., 2025). Nor does a fluent trace reveal what drove the answer: models biased toward "(A)" by few-shot examples pick it without mentioning the bias (Turpin et al., 2023).

Process labels also differ in what they measure. PRM800K asks humans whether each step is valid (Lightman et al., 2023), Math-Shepherd asks whether completions from a step still reach the answer (Section 3), and Setlur et al. (2024) measure how much a step changes the chance of eventual success. The three can assign different scores to the same step, yet all are called process rewards.

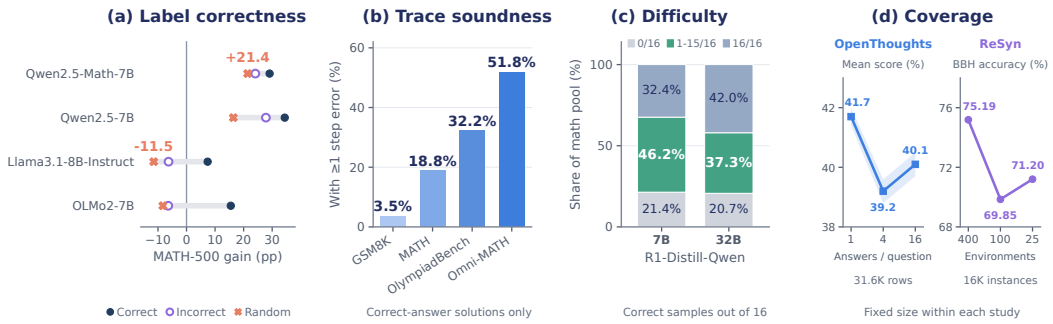

Figure 2: **Each property of useful data holds relative to something else.** (a) On MATH-500, random rewards lift Qwen but hurt Llama and OLMo (Shao et al., 2026). (b) Correct answers with a step error (Zheng et al., 2025). (c) Informative band shrinks at 32B (He et al., 2025a). (d) At fixed size, fewer environments hurt more than fewer questions (Guha et al., 2025; He et al., 2026).

There is conflicting evidence about whether such signals beat outcome rewards. Process supervision beats outcome supervision for choosing among candidates (Lightman et al., 2023), and progress rewards make online RL 5–6× more sample-efficient (Setlur et al., 2024). Yet DeepSeek-R1 found that a model-based PRM, while useful for reranking, invites reward hacking in large-scale RL (DeepSeek-AI et al., 2025), and Cheng et al. (2025) show that summing step rewards collapses RL from the start; scoring a solution by its weakest step keeps it stable, though PRM-only training still reward-hacks eventually. **A process score that ranks finished solutions well can still be gamed once a policy is trained against it**; how the score is used matters as much as what it measures.

4.3 Difficulty Depends on the Learner

Difficulty is not a property of a problem alone, but of how often the current model solves it. Bae et al. (2025) show that expected policy improvement is bounded below by the variance of per-problem success, so problems of intermediate difficulty teach most. DAPO applies this by dropping prompts whose samples are all correct or all wrong, which lifts AIME 2024 from 42 to 50 in its ablation (Yu et al., 2025).

The same pool therefore yields a different curriculum for each learner (Figure 2c). On Skywork-OR1's math pool, sampled 16 times per problem, R1-Distill-Qwen-32B solves 42.0% of problems on every sample versus 32.4% for the 7B model, leaving 37.3% and 46.2% informative (He et al., 2025a). Because this band moves as the learner improves, fixed orderings fall short: easy problems help early but must be faded out to avoid overfitting (Parashar et al., 2025), and curricula that track learning progress online generalize better to harder problems (Chen et al., 2025b).

There is conflicting evidence about problems the learner cannot yet solve. DAPO and Skywork-OR1 discard them because they produce no reward signal (Yu et al., 2025; He et al., 2025a), yet in on-policy distillation even a single hard problem is effective, and harder problems often yield larger gains (Hou et al., 2026). **Whether a problem is "too hard" depends on how it enters learning**, not only on the learner's pass rate.

4.4 More Rows Is Not More Coverage

A row count conflates new problems, new solutions to old problems, and repeated exposure. At a fixed 31.6K rows, OpenThoughts finds that 16 teacher answers for each of about 2K questions perform within 1.6 points of one answer for each of 31.6K questions on math, and better on science; it therefore multiplies its data at least sixteenfold by repeated sampling (Figure 2d; Guha et al., 2025). In a fixed-recipe multimodal benchmark, adding data beyond about 1,000 aligned examples leaves accuracy flat and mainly reduces run-to-run variance (Shin et al., 2026). Coverage can even shrink over generations: models trained recursively on generated data lose the tails of the original distribution (Shumailov et al., 2024).

There is conflicting evidence about whether diversity helps. ReSyn's ablations credit part of its gains to more kinds of reasoning environments: at a fixed 16K instances, 400 environments beat 100 by 5.3

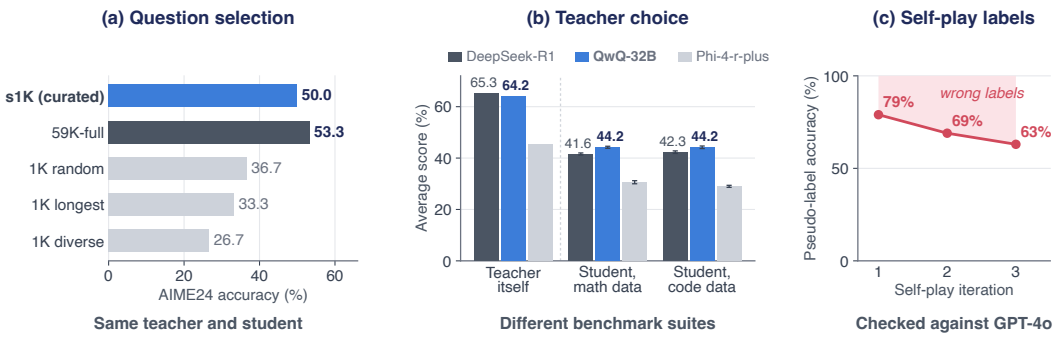

Figure 3: **Construction choices, measured.** (a) AIME24 accuracy of 1K questions chosen in different ways versus all 59K (Muennighoff et al., 2025). (b) QwQ-32B has the lower benchmark average yet trains the stronger student (Guha et al., 2025). (c) Agreement of R-Zero's self-generated labels with GPT-4o (Huang et al., 2026).

points on BBH (He et al., 2026). Sancaktar et al. (2026) similarly suggest that diversity and structure, rather than volume alone, become the limiting factor in scaling RL for code. Yet in multimodal reasoning, diversity heuristics and rewritten synthetic mixtures fail to beat a difficulty-filtered baseline (Shin et al., 2026). **Adding new kinds of tasks expands what a dataset covers, whereas reselecting or rewriting an existing pool mostly reweights it**; this helps reconcile the findings.

5 How Is Reasoning Data Built?

The properties in Section 4 are fixed by choices made while building data. We follow those choices in pipeline order: which questions are kept, who writes the traces, what happens to failures, whether the model writes its own tasks, how feedback is designed, and how training stages are chained. Figure 3 collects measured effects of three of these choices, Figure 4 shows where three others act, and Appendix D.3 lists complete systems.

5.1 Choosing Questions versus Filtering Answers

Several releases have asked how far a small, carefully chosen set of questions can replace a large pool. Muennighoff et al. (2025) select 1,000 of 59,000 questions by difficulty, diversity and quality and come close to training on all of them with 7 rather than 394 H100 GPU hours, whereas 1,000 random questions fall well short (Figure 3a). Ye et al. (2025) report that a small curated set outperforms models trained on 100 times more data. Guha et al. (2025) compare selection strategies directly: filtering questions by LLM-judged difficulty or response length beats embedding-based filters, but none of their answer-filtering methods yields significant gains. Together, these results suggest that **choosing which questions to train on matters more than filtering the answers to them**, at least when a strong teacher writes those answers.

5.2 The Best Solver Is Not Always the Best Teacher

There is conflicting evidence about whether a stronger model makes a better trace writer. Muennighoff et al. (2025) keep s1's questions and student fixed and find that replacing Gemini traces with DeepSeek-R1 traces nearly doubles AIME 2025 accuracy, from 26.7 to 50.0. Guha et al. (2025), on the other hand, find that QwQ-32B teaches better than DeepSeek-R1, even though its overall benchmark average is lower (Figure 3b). Two studies that select data for a specific student suggest a reconciliation: Hu et al. (2026) show, across nine teachers, that the strongest teacher's correct answer is not necessarily the best supervision for a given student, and Just et al. (2025) find that scoring each step by whether it follows from its premises beats scoring whole solutions by fluency. Kim et al. (2026) add that correct traces can even hurt, cutting AIME 2024 accuracy from 55% to 20% when they suppress expressions of uncertainty. Taken together, these results suggest that **a good teacher writes traces the student can learn from, not merely traces that are correct**.

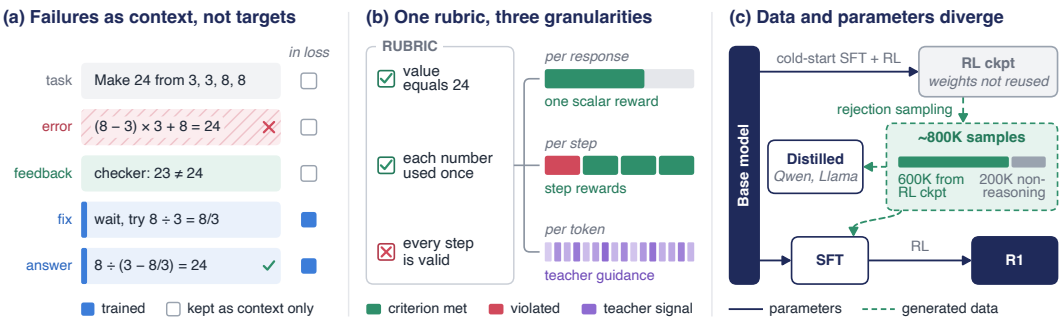

Figure 4: **Where construction choices act.** (a) Errors and checker output stay in context but are masked from the loss. (b) One rubric gives rewards per response, step or token; a single reward still credits the invalid step. (c) In DeepSeek-R1, data from an RL checkpoint train the original base and distill smaller models (DeepSeek-AI et al., 2025).

5.3 Failures Can Teach without Being Imitated

Several studies have asked what to do with failed attempts, which standard pipelines discard. Pan et al. (2025) fine-tune a software agent only on its successful trajectories but train a verifier on successes and failures alike, and combining the two raises resolve rates. Xu et al. (2025b) add previously discarded negative traces to distillation and match a model trained on about six times as much proprietary data. Logeswaran et al. (2026) score how far a web agent progressed toward its goal, which turns partially successful trajectories into usable training data. Other work keeps errors in the record without imitating them (Figure 4a): Chen et al. (2025c) insert reflections and corrections after error steps and mask the errors from the loss, Hao et al. (2026) roll back to the first harmful step and train on the teacher's correction, and Search-R1 masks retrieved text from the loss during RL (Jin et al., 2025). These results suggest that **failures are informative as negatives, as partial credit or as context, as long as they are not imitated**.

5.4 Self-Play Still Needs Curation

Several systems now let a model write its own training tasks, and they differ mainly in what decides which answers count as correct. Absolute Zero uses a code executor to validate both tasks and answers (Zhao et al., 2025a), and STP checks proofs in Lean, doubling the previous best on LeanWorkbook (Dong & Ma, 2025). SPICE grounds generated questions in a document corpus and reports that ungrounded self-play gains less (Liu et al., 2025a). Systems that rely on the model's own majority vote fare worse: in R-Zero, the accuracy of pseudo-labels, judged against GPT-4o, falls from 79% to 63% within three iterations (Huang et al., 2026) (Figure 3c), and Pan et al. (2026) show that maximizing self-consistency collapses output diversity until the model confidently reinforces its own errors. Even with a sound check, the generated questions can lose diversity after a few iterations (Mishra, 2026). **Self-play thus moves curation into the check that decides what counts as correct, rather than removing it**.

5.5 The Same Rubric Can Train Different Models

Where answers cannot be checked automatically, rubrics, lists of criteria a good response should meet, are a common source of feedback, and studies differ in where the rubric enters learning (Figure 4b). Gunjal et al. (2025) aggregate rubric scores into one reward and beat direct Likert-style judging by up to 31% on HealthBench. Xie et al. (2026) observe that such response-level scores reward many wrong steps and penalize many correct ones, and instead assign each rubric item to a step, which improves math accuracy. Gu et al. (2026) go further: the policy, shown the rubric, acts as its own token-level teacher. Others learn the rubric itself: Rubric-ARM jointly trains a rubric generator and a judge (Xu et al., 2026), and OnlineRubrics revises criteria during training to counter reward hacking (Rezaei et al., 2025). Fiedler (2026) caution that judge-based comparisons can mislead when calibration is shared across models. **A rubric is thus constructed data in its own right**: the same criteria can train a judge or a policy, at the level of responses, steps or tokens.

5.6 Stages That Help Alone Can Hurt Together

Most frontier recipes chain several stages, and data can travel along a different path from parameters. In DeepSeek-R1, about 600K correct outputs from a first-round RL checkpoint, together with about 200K non-reasoning samples, train the original base model before a second round of RL, and the same 800K samples also distill smaller models through SFT (DeepSeek-AI et al., 2025) (Figure 4c).

There is conflicting evidence about how supervised and RL stages combine. In Phi-4-reasoning, a short RL phase amplifies the gains of careful SFT (Abdin et al., 2025), and Ma et al. (2025) show that interleaving SFT on questions RL cannot solve beats either alone with far fewer demonstrations. Zhu et al. (2026), in contrast, find that sequential SFT on several tasks causes severe interference, whereas RL on the same tasks does not. Most reports change several stages at once, and gains often hinge on unclear comparisons (Hochlehnert et al., 2025); still, **supervised stages seem to help when they supply what RL cannot reach, and to interfere when they are stacked across tasks.**

6 What Does Scaling Evidence Establish?

Reasoning datasets now span several orders of magnitude, from millions of questions (Yuan et al., 2025a; Moshkov et al., 2025) down to a single prompt (Wang et al., 2025; Hou et al., 2026), and both ends report strong results. This section asks what "more data" counts, what it is measured against, and whether the test itself stays the same.

6.1 Few Prompts Do Not Mean Little Supervision

In general, more data helps: OpenThoughts reports significant gains as it scales its pipeline to 1.2 million examples (Guha et al., 2025). Studies that vary the number of distinct prompts, however, find that far fewer are needed than a large corpus suggests. Wang et al. (2025) show that RL on a single well-chosen example roughly doubles Qwen2.5-Math-1.5B's MATH500 accuracy. Hou et al. (2026) find that on-policy distillation from one prompt is also effective. Tan et al. (2026) report that when data are scarce, reusing high-quality problems works well. AR3PO matches DAPO at up to 4.2 times lower rollout cost, mostly by giving easy prompts fewer responses (Zhang et al., 2025a). These results suggest that **a small prompt pool is not a small supervision budget**: distinct prompts, sampled trajectories and supervised tokens are different resources, and counting only prompts hides which one was scarce (Figure 5).

6.2 Finding a Solution Is Not Selecting It

A model that writes 64 attempts at a hard problem, one of them correct, succeeds under pass@64, yet a user who gets one answer may not. Let A_k be the event that at least one of k candidates is correct and D_k the event that the answer picked from them is correct; then

$$P(D_k) = P(A_k)\,P(D_k \mid A_k), \qquad (1)$$

so delivered accuracy depends on both finding and selecting. Brown et al. (2024) show that coverage keeps rising with more samples, but selection without a verifier soon plateaus.

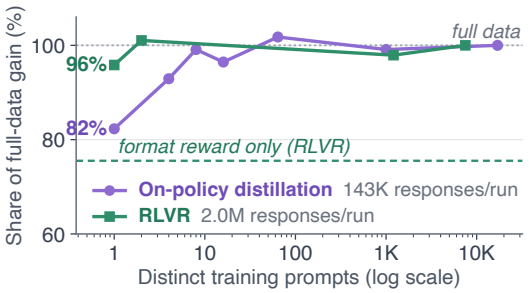

Figure 5: **Few prompts, most of the gain**, with sampled responses fixed within each study (Hou et al., 2026; Wang et al., 2025).

There is conflicting evidence about whether RL teaches new capabilities or makes existing ones more frequent (Table 2). Yue et al. (2025) find that RL-trained models win at small k but are overtaken by their base models at large k. Wu et al. (2026) likewise find that RL shrinks the set of reachable answers more than it expands it. Liu et al. (2025b), on the other hand, report that prolonged RL solves tasks on which the base model fails at any budget. Wen et al. (2025) find that RL extends the reasoning boundary under a metric that also checks the steps. Zhai et al. (2026) find that RL expands what tool-using agents can solve on compositional tasks. **Whether RL expands what a model can solve thus remains inconclusive**: the studies differ in base model, training length, metric and task type, and an answer not sampled at a finite budget is not thereby impossible.

6.3 Some Gains Come from Changing the Test

A recurring problem in empirical work is that reported gains may lie within the variation introduced by evaluation choices, and reasoning models are no exception. Hochlehnert et al. (2025) find that math reasoning scores shift with random seeds and decoding settings, and that many reported RL gains shrink under standardized evaluation. s1 raises accuracy simply by appending "Wait" whenever the model tries to stop thinking (Muennighoff et al., 2025). TTRL roughly triples AIME 2024 accuracy by training on the unlabeled test questions themselves (Zuo et al., 2025). Search agents find ground-truth answers online for about 3% of benchmark questions (Han et al., 2025). **A higher score can therefore reflect better data, better use of existing capability, or a changed test.**

Table 2: **Does RL expand what a model can solve?** ▼ narrows, ▲ expands; the studies differ in model, sample budget k and task.

Study	Model; budget	Finding
▼ Yue et al., 2025	Qwen2.5, Llama; $k \leq 1024$	base overtakes RL at large k
▼ Wu et al., 2026	ProRL-1.5B; $k \leq 8192$	more lost than gained (163 vs. 36)
▲ Liu et al., 2025b	R1-Distill-1.5B; $k \leq 256$	solves puzzles the base never does
▲ Wen et al., 2025	DAPO-Qwen-32B; $k \leq 1024$	ahead when steps are also checked
▲ Zhai et al., 2026	Qwen2.5-7B agent; $k \leq 64$	$+4$ of 100 questions

7 What Do the Disagreements Reveal?

One puzzle recurs throughout this primer. Random rewards lift Qwen2.5-Math-7B by 21 points on MATH-500 yet hurt Llama and OLMo (Shao et al., 2026). Alone, it looks like a paradox; alongside other findings, a pattern emerges. RL pipelines drop problems they never solve (Yu et al., 2025), whereas under on-policy distillation harder problems often teach the most (Hou et al., 2026), and the strongest teacher does not always teach best (Hu et al., 2026). In each case the questions stay the same; what changes is the learner, or the route by which data reach it: imitation, RL or distillation. **Much of what looks like disagreement about reasoning data is disagreement about who learns and how**; six of the seven conflicts in Appendix B.2 hinge mainly on learner or route.

A second puzzle concerns feedback. Corrupting up to 15% of rewards at random barely hurts RL (Plesner et al., 2026), yet one systematic flaw can sink it: a policy trained against an LLM judge learned to answer with little more than "Thought process:", and the judge kept rewarding it (Zhao et al., 2025b). Process reward models invite the same exploitation (DeepSeek-AI et al., 2025), and majority votes end up reinforcing a model's own mistakes (Pan et al., 2026). **Random errors seem to average out; errors the learner can exploit do not.**

Three controlled comparisons could test these patterns. **Teacher rankings:** crossing several teachers with several students on the same questions and budget tests whether rankings reorder, as s1 and OpenThoughts suggest (Muennighoff et al., 2025; Guha et al., 2025); **Learning route:** DAPO and on-policy distillation already share DAPO-Math-17K (Yu et al., 2025; Hou et al., 2026), so training one student on it both ways tests whether the most useful difficulty shifts with the route; **Error type:** injecting random and exploitable errors into one checker at matched rates tests whether exploitability, rather than frequency, harms training. If rankings and the most useful difficulty stay fixed across students and routes, or both kinds of error do equal harm, usefulness would belong to the data after all. Each test needs reports that state the learner, target and update rule, as in Appendix B.3 (Gebru et al., 2021; Bender & Friedman, 2018; Longpre et al., 2024; Hochlehnert et al., 2025).

8 Conclusion

Public reasoning data have grown from a thousand curated examples to over a million generated traces, and their feedback from answer checks to rubrics and interactive environments. Across this growth, the evidence reviewed here points to a simple lesson: a correct answer is not the same as a good lesson, and what a record teaches depends on who learns from it and how. The open challenge is to measure this dependence before training rather than discover it afterwards. As releases keep multiplying, we hope this primer helps the community ask of each new dataset not only what it contains, but whom it teaches and how.

AI USE STATEMENT

An AI assistant supported English drafting, translation checks, targeted literature checks, and LaTeX revision. The conceptual illustrations were generated with an image-generation model and checked against the relationships described in the text. The data charts redraw cited numerical values rather than new measurements. No new training experiments or empirical results were generated. The authors remain responsible for source selection, verification, and the final claims and wording.

ETHICS STATEMENT

This work releases no new data or models. The practices reviewed still involve privacy-bearing traces, expert or model judges, safety boundaries, agent actions, and costly training (Green et al., 2025; Arora et al., 2025; Jiang et al., 2025; Trivedi et al., 2024; DeepSeek-AI et al., 2025). Responsible releases require license, provenance, welfare, verifier-gaming, contamination and compute disclosures.

REPRODUCIBILITY STATEMENT

This primer reports no new experiments; every result it discusses is taken from a cited public source. Appendix A.1 describes how sources were found and selected, and Appendix A.2 how each reported number was checked against its source. Appendix B.1 lists the original numbers behind the rounded figures in the main text, together with what each study held fixed and what it changed, and Appendix B.2 pairs the conflicting findings discussed in Section 7. The search records are kept in the companion repository.

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025. doi: 10.48550/arXiv.2504.21318. URL https://arxiv.org/abs/2504.21318.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025. doi: 10.48550/arXiv.2502.17387. URL https://arxiv.org/abs/2502.17387.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi: 10.48550/arXiv.2505.08775. URL https://arxiv.org/abs/2505.08775.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023. doi: 10.48550/arXiv.2302.12433. URL https://arxiv.org/abs/2302.12433.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. *arXiv preprint arXiv:2504.03380*, 2025.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. doi: 10.48550/arXiv.2204.05862. URL https://arxiv.org/abs/2204.05862.

Kshitij Bansal, Sarah M. Loos, Markus N. Rabe, Christian Szegedy, and Stewart Wilcox. Holist: An environment for machine learning of higher order logic theorem proving. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pp. 454–463. PMLR, 2019. URL http://proceedings.mlr.press/v97/bansal19a.html.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. doi: 10.48550/arXiv.2407.21787. URL https://arxiv.org/abs/2407.21787.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/arXiv.2107.03374. URL https://arxiv.org/abs/2107.03374.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025a.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for LLM reasoning. *arXiv preprint arXiv:2505.14970*, 2025b.

Yihan Chen, Benfeng Xu, Xiaorui Wang, Yongdong Zhang, and Zhendong Mao. Training LLM-based agents with synthetic self-reflected trajectories and partial masking. *arXiv preprint arXiv:2505.20023*, 2025c.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2022. doi: 10.48550/arXiv.2109.00122. URL https://arxiv.org/abs/2109.00122.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *arXiv preprint arXiv:2504.15275*, 2025.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang,

Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2025. doi: 10.48550/arXiv.2412.05467. URL https://arxiv.org/abs/2412.05467.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2024. doi: 10.48550/arXiv.2310.01377. URL https://arxiv.org/abs/2310.01377.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, Online, jun 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.365. URL https://aclanthology.org/2021.naacl-main.365/.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.1038/s41586-025-09422-z. URL https://arxiv.org/abs/2501.12948.

Kefan Dong and Tengyu Ma. STP: Self-play LLM theorem provers with iterative conjecturing and proving. *arXiv preprint arXiv:2502.00212*, 2025.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge

work tasks? *arXiv preprint arXiv:2403.07718*, 2024. doi: 10.48550/arXiv.2403.07718. URL https://arxiv.org/abs/2403.07718.

James Fiedler. Bias and uncertainty in LLM-as-a-judge estimation. *arXiv preprint arXiv:2605.06939*, 2026.

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. doi: 10.48550/arXiv.2503.01307. URL https://arxiv.org/abs/2503.01307.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022. doi: 10.48550/arXiv.2210.10760. URL https://arxiv.org/abs/2210.10760.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2021. doi: 10.48550/arXiv.1803.09010. URL https://arxiv.org/abs/1803.09010.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*, 2025. doi: 10.48550/arXiv.2411.04872. URL https://arxiv.org/abs/2411.04872.

Tommaso Green, Martin Gubri, Haritz Puerto, Sangdoo Yun, and Seong Joon Oh. Leaky thoughts: Large reasoning models are not private thinkers. *arXiv preprint arXiv:2506.15674*, 2025. doi: 10.48550/arXiv.2506.15674. URL https://arxiv.org/abs/2506.15674.

Siyi Gu, Jialin Chen, Sophia Zhou, Arman Cohan, and Rex Ying. Rethinking reward supervision: Rubric-conditioned self-distillation. *arXiv preprint arXiv:2606.19327*, 2026.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025. doi: 10.48550/arXiv.2506.04178. URL https://arxiv.org/abs/2506.04178.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*, 2023. doi: 10.48550/arXiv.2308.11462. URL https://arxiv.org/abs/2308.11462.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025. doi: 10.48550/arXiv.2507.17746. URL https://arxiv.org/abs/2507.17746.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *arXiv preprint arXiv:2406.18495*, 2024. doi: 10.48550/arXiv.2406.18495. URL https://arxiv.org/abs/2406.18495.

Ziwen Han, Meher Mankikar, Julian Michael, and Zifan Wang. Search-time data contamination. *arXiv preprint arXiv:2508.13180*, 2025. doi: 10.48550/arXiv.2508.13180. URL https://arxiv.org/abs/2508.13180.

Longkun Hao, Hongyu Lin, Hao Li, Zhuowen Liu, Zhichao Yang, Haojie Hao, Dongshuo Huang, Haitao Yang, Hongyu Ge, Ming jie Xie, Yanjun Wu, Zi Hao Yin, Yan Bai, and Yihang Lou. Speculative rollback correction for quality-diverse web agent imitation. *arXiv preprint arXiv:2606.12485*, 2026.

Andre He, Nathaniel Weir, Kaj Bostrom, Allen Nie, Darion Cassel, Sam Bayless, and Huzefa Rangwala. ReSyn: Autonomously scaling synthetic environments for reasoning models. *arXiv preprint arXiv:2602.20117*, 2026.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025a. doi: 10.48550/arXiv.2505.22312. URL https://arxiv.org/abs/2505.22312.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*, 2025b. doi: 10.48550/arXiv.2504.11456. URL https://arxiv.org/abs/2504.11456.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*, 2021a. doi: 10.48550/arXiv.2105.09938. URL https://arxiv.org/abs/2105.09938.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021b. doi: 10.48550/arXiv.2103.06268. URL https://arxiv.org/abs/2103.06268.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021c. doi: 10.48550/arXiv.2103.03874. URL https://arxiv.org/abs/2103.03874.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. *arXiv preprint arXiv:2504.07086*, 2025. doi: 10.48550/arXiv.2504.07086. URL https://arxiv.org/abs/2504.07086.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025. doi: 10.48550/arXiv.2503.24290. URL https://arxiv.org/abs/2503.24290.

Zhengyu Hu, Zheyuan Xiao, Linxin Song, Fengqing Jiang, Yuetai Li, Zhihan Xiong, Yue Liu, Junhao Lin, Yao Su, Lijie Hu, Kaize Ding, Teng Xiao, and Radha Poovendran. When the strongest teacher is not the best teacher: Student-centric answer selection. *arXiv preprint arXiv:2605.26872*, 2026.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2026. doi: 10.48550/arXiv.2508.05004. URL https://arxiv.org/abs/2508.05004.

Yuzhen Huang, Weihao Zeng, Xingshan Zeng, Qi Zhu, and Junxian He. From accuracy to robustness: A study of rule- and model-based verifiers in mathematical reasoning. *arXiv preprint arXiv:2505.22203*, 2025. doi: 10.48550/arXiv.2505.22203. URL https://arxiv.org/abs/2505.22203.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

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. doi: 10.48550/arXiv.2403.07974. URL https://arxiv.org/abs/2403.07974.

Naman Jain, Jaskirat Singh, Manish Shetty, Liang Zheng, Koushik Sen, and Ion Stoica. R2e-gym: Procedural environments and hybrid verifiers for scaling open-weights swe agents. *arXiv preprint arXiv:2504.07164*, 2025. doi: 10.48550/arXiv.2504.07164. URL https://arxiv.org/abs/2504.07164.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities. *arXiv preprint arXiv:2502.12025*, 2025. doi: 10.48550/arXiv.2502.12025. URL https://arxiv.org/abs/2502.12025.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019. doi: 10.48550/arXiv.1909.06146. URL https://arxiv.org/abs/1909.06146.

Hoang Anh Just, Myeongseob Ko, and Ruoxi Jia. The signal is in the steps: Local scoring for reasoning data selection. *arXiv preprint arXiv:2510.03988*, 2025.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025. doi: 10.48550/arXiv.2510.13786. URL https://arxiv.org/abs/2510.13786.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of llms? *arXiv preprint arXiv:2603.24472*, 2026. doi: 10.48550/arXiv.2603.24472. URL https://arxiv.org/abs/2603.24472.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024. doi: 10.48550/arXiv.2405.01535. URL https://arxiv.org/abs/2405.01535.

Masahiro Koreeda and Christopher Manning. Contractnli: A dataset for document-level natural language inference for contracts. In *EMNLP*, 2021.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. doi: 10.48550/arXiv.2403.13787. URL https://arxiv.org/abs/2403.13787.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2025. doi: 10.48550/arXiv.2411.15124. URL https://arxiv.org/abs/2411.15124.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702.

Joshua Ong Jun Leang, Giwon Hong, Wenda Li, and Shay B. Cohen. Theorem prover as a judge for synthetic data generation. *arXiv preprint arXiv:2502.13137*, 2025.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. LLMs can easily learn to reason from demonstrations: Structure, not content, is what matters! *arXiv preprint arXiv:2502.07374*, 2025.

Minghao Li, Feifan Song, Bowen Yu, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A benchmark for tool-augmented llms. *CoRR*, abs/2304.08244, 2023. doi: 10.48550/ARXIV.2304.08244. URL https://doi.org/10.48550/arXiv.2304.08244.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. doi: 10.48550/arXiv.2305.20050. URL https://arxiv.org/abs/2305.20050.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2022. doi: 10.48550/arXiv.2109.07958. URL https://arxiv.org/abs/2109.07958.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. SPICE: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025a.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025b. doi: 10.48550/arXiv.2505.24864. URL https://arxiv.org/abs/2505.24864.

Zuxin Liu, Thai Hoang, Jianguo Zhang, Ming Zhu, Tian Lan, Shirley Kokane, Juntao Tan, Weiran Yao, Zhiwei Liu, Yihao Feng, Rithesh Murthy, Liangwei Yang, Silvio Savarese, Juan Carlos Niebles, Huan Wang, Shelby Heinecke, and Caiming Xiong. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *arXiv preprint arXiv:2406.18518*, 2024. doi: 10.48550/arXiv.2406.18518. URL https://arxiv.org/abs/2406.18518.

Lajanugen Logeswaran, Jaekyeom Kim, Sungryull Sohn, Creighton Glasscock, and Honglak Lee. Scaling web agent training through automatic data generation and fine-grained evaluation. *arXiv preprint arXiv:2602.12544*, 2026.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6, 2024.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities. *arXiv preprint arXiv:2408.04682*, 2025. doi: 10.48550/arXiv.2408.04682. URL https://arxiv.org/abs/2408.04682.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024. doi: 10.48550/arXiv.2402.04249. URL https://arxiv.org/abs/2402.04249.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiguang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jiaren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Linge Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujie Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songquan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, Yonghong Duan, Yongxiang Fu, Yongyi Hu, Yu Gao, Yuanxiang Fan, Yufeng Yang, Yuhao Li, Yulin Hu, Yunan Huang, Yunji Li, Yunzhi Xu, Yuxin Mao, Yuxuan Shi, Yuze Wenren, Zehan Li, Zelin Li, Zhanxu Tian, Zhengmao Zhu, Zhenhua Fan, Zhenzhen Wu, Zhichao Xu, Zhihang Yu, Zhiheng Lyu, Zhuo Jiang, Zibo Gao, Zijia Wu, Zijian Song, and Zijun Sun. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025. doi: 10.48550/arXiv.2506.13585. URL https://arxiv.org/abs/2506.13585.

Vaibhav Mishra. Preventing curriculum collapse in self-evolving reasoning systems. *arXiv preprint arXiv:2603.13309*, 2026.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, Teven Le Scao, Yihan Wang, Adam Yang, Alexander H. Liu, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Andy Ehrenberg, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jean-Hadrien Chabran, Jean-Malo Delignon, Joachim Studnia, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Kush Jain, Lingxiao Zhao, Louis Martin, Luyu Gao, Lélio Renard Lavaud, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Maximilian Augustin, Mickaël Seznec, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patrick von Platen, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Romain Sauvestre, Rémi Delacourt, Sanchit Gandhi, Sandeep Subramanian, Shashwat Dalal, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, and Yunhao Tang. Magistral. *arXiv preprint arXiv:2506.10910*, 2025. doi: 10.48550/arXiv.2506.10910. URL https://arxiv.org/abs/2506.10910.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025. doi: 10.48550/arXiv.2504.16891. URL https://arxiv.org/abs/2504.16891.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. doi: 10.48550/arXiv.2501.19393. URL https://arxiv.org/abs/2501.19393.

OpenAI. Learning to reason with LLMs. Technical report, Technical report, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. doi: 10.48550/arXiv.2203.02155. URL https://arxiv.org/abs/2203.02155.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. Training software engineering agents and verifiers with swe-gym. *arXiv preprint arXiv:2412.21139*, 2025. doi: 10.48550/arXiv.2412.21139. URL https://arxiv.org/abs/2412.21139.

Teng Pan, Yuchen Yan, Zixuan Wang, Ruiqing Zhang, Guiyang Hou, Wenqi Zhang, Weiming Lu, Jun Xiao, and Yongliang Shen. Coverrl: Breaking the consensus trap in label-free reasoning via generator-verifier co-evolution. *arXiv preprint arXiv:2603.17775*, 2026. doi: 10.48550/arXiv.2603.17775. URL https://arxiv.org/abs/2603.17775.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and Shuiwang Ji. Curriculum reinforcement learning from easy to hard tasks improves LLM reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023. doi: 10.48550/arXiv.2305.15334. URL https://arxiv.org/abs/2305.15334.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://gorilla.cs.berkeley.edu/blogs/13_bfcl_v3_multi_turn.html.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster,

Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu,

Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao

Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Ben Wu, Jacek Karwowski, Davide Scaramuzza, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Steven Dillmann, Zhengxiang Wang, Junyu Luo, Hugo Lunn, Artem Gazizov, Haoran Qiu, Allen G Hart, Rickard Brüel Gabrielsson, Ido Akov, Artem Lukoianov, Summer Yue, Alexandr Wang, and Dan Hendrycks. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2026. doi: 10.1038/s41586-025-09962-4. URL https://arxiv.org/abs/2501.14249.

Andreas Plesner, Francisco Guzmán, and Anish Athalye. An imperfect verifier is good enough: Learning with noisy rewards. *arXiv preprint arXiv:2604.07666*, 2026. doi: 10.48550/arXiv.2604.07666. URL https://arxiv.org/abs/2604.07666.

Qwen Team. Qwen3-Coder: Agentic coding in the world, 2025.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*, 2024a. doi: 10.48550/arXiv.2406.02900. URL https://arxiv.org/abs/2406.02900.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2024b. doi: 10.48550/arXiv.2305.18290. URL https://arxiv.org/abs/2305.18290.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Toyama, Robert Berry, Divya Tyamagundlu, Timothy Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573*, 2025. doi: 10.48550/arXiv.2405.14573. URL https://arxiv.org/abs/2405.14573.

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. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*, 2025. doi: 10.48550/arXiv.2510.07284. URL https://arxiv.org/abs/2510.07284.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2024. doi: 10.48550/arXiv.2308.01263. URL https://arxiv.org/abs/2308.01263.

Cansu Sancaktar, David Zhang, Gabriel Synnaeve, and Taco Cohen. A deep dive into scaling RL for code generation with synthetic data and curricula. *arXiv preprint arXiv:2603.24202*, 2026.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2022. doi: 10.48550/arXiv.2110.08207. URL https://arxiv.org/abs/2110.08207.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. doi: 10.48550/arXiv.2302.04761. URL https://arxiv.org/abs/2302.04761.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*, 2024. doi: 10.48550/arXiv.2410.08146. URL https://arxiv.org/abs/2410.08146.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in rlvr. *arXiv preprint arXiv:2506.10947*, 2026. doi: 10.48550/arXiv.2506.10947. URL https://arxiv.org/abs/2506.10947.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Yosub Shin, Michael Buriek, Boris Sobolev, Pavel Bushuyeu, Vikas Kumar, Haoyang Xu, Samuel Watson, and Igor Molybog. Data-efficient curation for multimodal reasoning under fixed training protocols. *arXiv preprint arXiv:2601.10922*, 2026.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07566-y. URL http://dx.doi.org/10.1038/s41586-024-07566-y.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2024. doi: 10.48550/arXiv.2312.06585. URL https://arxiv.org/abs/2312.06585.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. doi: 10.48550/arXiv.2212.13138. URL https://arxiv.org/abs/2212.13138.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024. doi: 10.48550/arXiv.2409.13740. URL https://arxiv.org/abs/2409.13740.

Zelin Tan, Hejia Geng, Xiaohang Yu, Mulei Zhang, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, Philip Torr, and Lei Bai. Scaling behaviors of llm reinforcement learning post-training: An empirical study in mathematical reasoning. *arXiv preprint arXiv:2509.25300*, 2026. doi: 10.48550/arXiv.2509.25300. URL https://arxiv.org/abs/2509.25300.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025. doi: 10.48550/arXiv.2501.12599. URL https://arxiv.org/abs/2501.12599.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. *arXiv preprint arXiv:2407.18901*, 2024. doi: 10.48550/arXiv.2407.18901. URL https://arxiv.org/abs/2407.18901.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28, 2015. doi: 10.1186/S12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023. doi: 10.48550/arXiv.2305.04388. URL https://arxiv.org/abs/2305.04388.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020. doi: 10.48550/arXiv.2004.14974. URL https://arxiv.org/abs/2004.14974.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2024. doi: 10.48550/arXiv.2312.08935. URL https://arxiv.org/abs/2312.08935.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025. doi: 10.48550/arXiv.2504.20571. URL https://arxiv.org/abs/2504.20571.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2023. doi: 10.48550/arXiv.2212.10560. URL https://arxiv.org/abs/2212.10560.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2022. doi: 10.48550/arXiv.2109.01652. URL https://arxiv.org/abs/2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2023. doi: 10.48550/arXiv.2201.11903. URL https://arxiv.org/abs/2201.11903.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025. doi: 10.48550/arXiv.2506.14245. URL https://arxiv.org/abs/2506.14245.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark. *arXiv preprint arXiv:2406.19314*, 2025. doi: 10.48550/arXiv.2406.19314. URL https://arxiv.org/abs/2406.19314.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2026. doi: 10.48550/arXiv.2507.14843. URL https://arxiv.org/abs/2507.14843.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*, 2025.

Weichu Xie, Haozhe Zhao, Wenpu Liu, Yongfu Zhu, Liang Chen, Minghao Ye, Zirong Chen, Yuqi Xu, Shuai Dong, Ziyue Wang, Xinbo Xu, Kean Shi, Ruoyu Wu, Xiaoying Zhang, Wenqi Shao, Baobao Chang, Nan Duan, and Jiaqi Wang. Step-wise rubric rewards for LLM reasoning. *arXiv preprint arXiv:2605.17291*, 2026.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024a. doi: 10.48550/arXiv.2405.14333. URL https://arxiv.org/abs/2405.14333.

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*, 2024b. doi: 10.48550/arXiv.2408.08152. URL https://arxiv.org/abs/2408.08152.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2025a. doi: 10.48550/arXiv.2412.14161. URL https://arxiv.org/abs/2412.14161.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026. doi: 10.48550/arXiv.2602.01511. URL https://arxiv.org/abs/2602.01511.

Shuyao Xu, Cheng Peng, Jiangxuan Long, Weidi Xu, Wei Chu, and Yuan Qi. Harnessing negative signals: Reinforcement distillation from teacher data for LLM reasoning. *arXiv preprint arXiv:2505.24850*, 2025b.

Zhangchen Xu, Yuetai Li, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. Tinyv: Reducing false negatives in verification improves rl for llm reasoning. *arXiv preprint arXiv:2505.14625*, 2025c. doi: 10.48550/arXiv.2505.14625. URL https://arxiv.org/abs/2505.14625.

Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. *arXiv preprint arXiv:2503.02951*, 2025d. doi: 10.48550/arXiv.2503.02951. URL https://arxiv.org/abs/2503.02951.

Lecheng Yan, Ruizhe Li, Guanhua Chen, Qing Li, Jiahui Geng, Wenxi Li, Vincent Wang, and Chris Lee. Spurious rewards paradox: Mechanistically understanding how rlvr activates memorization shortcuts in llms. *arXiv preprint arXiv:2601.11061*, 2026. doi: 10.48550/arXiv.2601.11061. URL https://arxiv.org/abs/2601.11061.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J. Prenger, and Animashree Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/4441469427094f8873d0fecb0c4e1cee-Abstract-Datasets_and_Benchmarks.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023. doi: 10.48550/arXiv.2210.03629. URL https://arxiv.org/abs/2210.03629.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025. doi: 10.48550/arXiv.2502.03387. URL https://arxiv.org/abs/2502.03387.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. doi: 10.48550/arXiv.2503.14476. URL https://arxiv.org/abs/2503.14476.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Ilia Kulikov, Kyunghyun Cho, Dong Wang, Yuandong Tian, Jason E Weston, and Xian Li. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *arXiv preprint arXiv:2502.13124*, 2025a. doi: 10.48550/arXiv.2502.13124. URL https://arxiv.org/abs/2502.13124.

Youliang Yuan, Qiuyang Mang, Jingbang Chen, Hong Wan, Xiaoyuan Liu, Junjielong Xu, Jen tse Huang, Wenxuan Wang, Wenxiang Jiao, and Pinjia He. Curing miracle steps in LLM mathematical reasoning with rubric rewards. *arXiv preprint arXiv:2510.07774*, 2025b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025. doi: 10.48550/arXiv.2504.13837. URL https://arxiv.org/abs/2504.13837.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. *arXiv preprint arXiv:2203.14465*, 2022. doi: 10.48550/arXiv.2203.14465. URL https://arxiv.org/abs/2203.14465.

Zhiyuan Zhai, Wenjing Yan, Xiaodan Shao, and Xin Wang. Does rl expand the capability boundary of llm agents? a pass@(k,t) analysis. *arXiv preprint arXiv:2604.14877*, 2026. doi: 10.48550/arXiv.2604.14877. URL https://arxiv.org/abs/2604.14877.

Yuheng Zhang, Wenlin Yao, Changlong Yu, Yao Liu, Qingyu Yin, Bing Yin, Hyokun Yun, and Lihong Li. Improving sampling efficiency in RLVR through adaptive rollout and response reuse. *arXiv preprint arXiv:2509.25808*, 2025a.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025b. doi: 10.48550/arXiv.2501.07301. URL https://arxiv.org/abs/2501.07301.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025a. doi: 10.48550/arXiv.2505.03335. URL https://arxiv.org/abs/2505.03335.

Bingchen Zhao, Dhruv Srikanth, Yuxiang Wu, and Zhengyao Jiang. SpecBench: Measuring reward hacking in long-horizon coding agents. *arXiv preprint arXiv:2605.21384*, 2026.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025b. doi: 10.48550/arXiv.2507.08794. URL https://arxiv.org/abs/2507.08794.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, Vienna, Austria, jul 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/. arXiv:2412.06559.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=9ZPegFuFTFv.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. doi: 10.48550/arXiv.2306.05685. URL https://arxiv.org/abs/2306.05685.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help?: assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pp. 159–168. ACM, June 2021. doi: 10.1145/3462757.3466088. URL http://dx.doi.org/10.1145/3462757.3466088.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*, 2021. doi: 10.48550/arXiv.2105.07624. URL https://arxiv.org/abs/2105.07624.

Kejian Zhu, Zhuoran Jin, Shangqing Tu, Hongbang Yuan, Yushi Bai, Kang Liu, Juanzi Li, and Jun Zhao. SFT conflicts, RL coexists: A theoretical and empirical analysis of multi-task learning for LLMs. *arXiv preprint arXiv:2608.03573*, 2026.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2025. doi: 10.48550/arXiv.2310.17631. URL https://arxiv.org/abs/2310.17631.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, Thong Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, Binyuan Hui, Niklas Muennighoff, David Lo, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*, 2025. doi: 10.48550/arXiv.2406.15877. URL https://arxiv.org/abs/2406.15877.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025. doi: 10.48550/arXiv.2504.16084. URL https://arxiv.org/abs/2504.16084.

How the appendix is organized. Appendix A explains where the evidence comes from and how far it reaches. Appendix B reports what the studies behind Sections 4–7 found: exact numbers behind single-comparison claims, the seven conflicts with two filled reporting cards, and the evidence on checker errors, correct traces, scaling and RL capability. Appendix C shows how the mechanisms of Sections 2 and 4 play out on constructed examples. Appendix D collects reference material: a glossary, the public releases of Section 3, and complete training systems. Numbers attached to a citation are as reported by the source; numbers we derive for constructed examples, in Appendices B.4, B.7 and C, are illustrative.

A How the Evidence Was Collected and Its Limits

This section documents the evidence base introduced in Section 2: how it was gathered and checked (A.1–A.2), what it covers beyond the main text (A.3), and where it is thin (A.4).

A.1 Search and Inclusion

The primer cites 156 works published between 2015 and September 2026: 25 before 2023, 15 in 2023, 25 in 2024, 67 in 2025 and 24 in 2026. The search was iterative rather than a fixed systematic protocol: starting from widely used reasoning datasets, systems and benchmarks, we followed their references and the work citing them; the search records are kept in the companion repository. We included work that releases or constructs reasoning data, supplies the checkers and benchmarks such data rely on, or tests how data choices change training, across the domains listed in Section 2; earlier instruction and preference data enter as precedents. Preprints and technical reports are included because much of this area appears first in that form. Sources play four roles: 92 are discussed in the main text, 3 appear only elsewhere in the appendix, 55 appear only in the coverage clusters of Appendix A.3, and 6 are complete systems listed only in Appendix D.3.

A.2 How Numbers Were Checked

Every number we report was checked against its source: AI assistance located candidate passages, and the authors confirmed each number in the paper itself. Numbers are taken as reported, not rerun, from the latest arXiv version as of September 2026, and only from a paper's text or tables, not from figures; where the text and tables disagree, we follow the tables. Differences between scores are given in points and relative changes as percentages or multiples; where a paper reports a percentage without saying whether it is absolute or relative, we recompute it from the underlying scores. Rounded comparisons such as "nearly doubles" are backed by exact numbers in Appendix B.1.

A.3 What the Bibliography Covers beyond the Main Text

Beyond the sources the main text cites for its arguments, the bibliography covers eight clusters of benchmarks, datasets and audits that a reader should know exist (Table A.1). These entries show the extent of the survey; none is cited here as evidence for a claim.

Cluster	What it adds	n	Representative sources
Live evaluations	fresh or very hard tests	4	Phan et al., 2026; Rein et al., 2023; Glazer et al., 2025; White et al., 2025
Math, code tests	answers a program checks	6	Cobbe et al., 2021; Hendrycks et al., 2021c; Chen et al., 2021; Hendrycks et al., 2021a; Zhuo et al., 2025; Jain et al., 2024
Formal proofs	machine-checked proofs	6	Zheng et al., 2022; Yang et al., 2023; Azerbayev et al., 2023; Bansal et al., 2019; Xin et al., 2024a;b
Instruction data	post-training precedents	7	Wang et al., 2023; Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Cui et al., 2024; Bai et al., 2022; Rafailov et al., 2024b
Tool use, agents	environment feedback	12	Yao et al., 2023; Schick et al., 2023; Li et al., 2023; Patil et al., 2023; Liu et al., 2024; Patil et al., 2025; Lu et al., 2025; Chezelles et al., 2025; Zhou et al., 2024; Drouin et al., 2024; Rawles et al., 2025; Xu et al., 2025a
Science, biomedicine	answers judged on evidence	6	Dasigi et al., 2021; Wadden et al., 2020; Skarlinski et al., 2024; Singhal et al., 2022; Jin et al., 2019; Tsatsaronis et al., 2015
Finance, law	expert-annotated tasks	7	Islam et al., 2023; Chen et al., 2022; Zhu et al., 2021; Guha et al., 2023; Zheng et al., 2021; Hendrycks et al., 2021b; Koreeda & Manning, 2021
Judges, safety	judge and reward failures	9	Lin et al., 2022; Mazeika et al., 2024; Han et al., 2024; Röttger et al., 2024; Lambert et al., 2024; Zhu et al., 2025; Kim et al., 2024; Gao et al., 2022; Rafailov et al., 2024a

Table A.1: **Coverage clusters.** Representative sources, mostly beyond those discussed in the main text; 55 of the 57 are cited nowhere else. n: number of entries listed.

A.4 Limitations

The evidence behind this primer is uneven across domains and model families, and each finding inherits the conditions of its source: numbers are taken as reported (Appendix A.2), and we did not audit contamination or validate checkers. The search may have missed relevant work (Appendix A.1),

and 91 of the 156 sources appeared in 2025 or 2026, many as preprints that have not yet been peer reviewed. The evidence is densest in mathematics, and cross-domain examples do not show that effects carry over uniformly. It is also concentrated in one model family: 11 of the 13 comparisons in Table B.1 that train a model use a Qwen learner (Table B.2), and Qwen's pretraining appears to prime it for reasoning in ways other families do not share (Shao et al., 2026). This limit matters most for our own thesis, since the claim that usefulness depends on the learner rests on only a few learners. Most results come from a single training run, and scores on small benchmarks shift with seeds and decoding settings (Hochlehnert et al., 2025). Finally, the labels in Table B.3 are our judgment, studies that find no effect may be less likely to be released, and complete systems report only parts of their pipelines (Table D.2).

Within these limits, the central pattern does not rest on any single study: by our labels in Table B.3, six of the seven conflicts differ mainly in learner or route, and the cleanest instance compares two learners inside one study (Shao et al., 2026). Testing the pattern further requires the matched comparisons proposed in Section 7, run on model families beyond Qwen and reported with the fields of the reporting cards in Appendix B.3.

B Evidence Behind the Lessons

This section gives the evidence behind the lessons of Sections 4–7. Appendices B.1–B.3 cover the comparisons themselves: the exact numbers behind claims that rest on a single comparison, the two sides of each conflict, and two filled reporting cards. Appendices B.4–B.7 expand four lessons: how checker errors affect training, when correct traces teach less, what scaling results count, and whether RL adds capability. Numbers follow the rules of Appendix A.2.

B.1 Exact Numbers behind Single-Comparison Claims

Table B.1 gives the exact numbers behind the rounded figures in the main text and behind each bold conclusion of Sections 4–6 that rests on a single comparison; conflicts are in Appendix B.2. Each row states what was held fixed and what changed. Five rows need more than one number. The 1K selection of s1 also cuts training from 394 to 7 GPU-hours (Muennighoff et al., 2025). The OpenThoughts change lowers the eight-benchmark average by 1.6 points on math data but raises it by 1.2 on science data (Guha et al., 2025). More examples in Shin et al. (2026) narrow the spread across runs rather than raise the mean. AR3PO reaches DAPO's average with up to 4.2 times fewer sampled responses (Zhang et al., 2025a). The main text's "up to 31%" for RaR is a relative gain from its main experiment on the full HealthBench, whose scores appear only in a figure, so the row reports the HealthBench-1k ablation (Gunjal et al., 2025).

Table B.2 adds the learner, the test and the scoring behind each row. On AIME's 30 problems, one problem moves a score by 3.3 points, and a single score near 50% has a binomial standard error of about 9 points. Comparing two scores needs the standard error of their difference: 11 to 13 points if the two scores are treated as independent, and less if they are paired on the same problems, which would require per-problem results that the papers do not report. Under the independent approximation, only the TTRL gap among the AIME rows exceeds two standard errors (27.3 points against a standard error of 10.8), and the Gemini $\rightarrow$ R1 gap is about two (23.3 against 12.2). The "Wait", 59K $\rightarrow$ 1K and selected $\rightarrow$ random gaps (6.7, 3.3 and 13.3 points) cannot be separated from question-sampling noise without per-problem results. On MATH500's 500 problems, the standard error of a difference is about 3 points, far below the 37.6-point gain from RL on 1 example. Seeds add further noise (Hochlehnert et al., 2025), and only Shin et al. (2026) repeat training.

B.2 Seven Conflicts, Mostly Separated by Learner or Route

By our labels, six of the seven conflicts flagged in Sections 4–6 differ mainly in the learner or the route, as Section 7 argues; Table B.3 pairs one finding from each side. We mark *learner* when the learner is the most salient difference, and *route* when the route is. A route is the recipe by which a record reaches the learner, such as SFT, rejection sampling, RLVR, on-policy distillation (OPD) or selection only (best-of-N); choosing a route fixes the target and the update rule (Appendix C.1). We mark *intervention* when the two sides change the data in different ways. These labels are our judgment: a dot marks the most salient difference, not the only one.

§	Held fixed	Before	Changed	After	Metric
4.2a	annotation protocol Zheng et al., 2025	3.5	GSM8K → Omni-MATH	51.8	wrong steps (%)
4.2b	policy, questions Setlur et al., 2024	1×	outcome → progress reward	5–6×	sample efficiency
4.4a	31.6K rows, teacher Guha et al., 2025	41.7	1 → 16 answers each	40.1	8-task avg.
4.4b	model, settings Shin et al., 2026	0.475	1K → 7.5K examples	0.469	suite score
5.1a	model, settings Muennighoff et al., 2025	53.3	59K → 1K selected	50.0	AIME24
5.1b	model, settings Muennighoff et al., 2025	50.0	selected → random 1K	36.7	AIME24
5.2	prompts, model Muennighoff et al., 2025	26.7	Gemini → R1 traces	50.0	AIME25
5.3	base model Xu et al., 2025b	83.2	800K closed → 131K open‡	83.1	MATH500
5.4	LeanWorkbook Dong & Ma, 2025	13.2	expert iteration → STP‡	28.5	proved (%)
5.5	policy, GRPO, judge Gunjal et al., 2025	23.9	Likert → rubric-guided judge	35.9	HealthBench-1k
5.6	model Ma et al., 2025	49.0	SFT 46K → interleaved 8.6K	52.6	6-task avg.
6.1a	Qwen2.5-Math-1.5B Wang et al., 2025	36.0	base → RL on 1 example	73.6	MATH500
6.1b	Qwen2.5-7B Zhang et al., 2025a	42.9	DAPO → AR3PO sampling	43.1	avg.
6.3a	s1-32B Muennighoff et al., 2025	50.0	append “Wait”	56.7	AIME24
6.3b	Qwen2.5-Math-7B Zuo et al., 2025	12.9	base → TTRL on test questions	40.2	AIME24

Table B.1: **Controlled comparisons behind the main text.** §: main-text section that uses the row; a letter distinguishes rows from the same section, and Table B.2 uses the same keys. ‡ Compares against a different system, so less is held fixed.

§	Learner	Test (items)	Scoring	Difference
4.2a	– (annotation study)	correct solutions (200; 500)	expert labels	7 → 259 flawed
4.2b	Gemma 2B, 9B	MATH (not stated)	not stated	–
4.4a	Qwen2.5-7B-Inst.	8 benchmarks	repeated evaluation	−1.6 points
4.4b	Qwen2.5-VL-7B	10 benchmarks (16,640)	mean of 5 vs. 3 runs	−0.006 (suite score)
5.1a	Qwen2.5-32B-Inst.	AIME24 (30)	greedy, 30K-token cap	16 → 15 problems
5.1b	Qwen2.5-32B-Inst.	AIME24 (30)	greedy, 30K-token cap	15 → 11 problems
5.2	Qwen2.5-32B-Inst.	AIME25 (30)	greedy	8 → 15 problems
5.3	Qwen2.5-Math-1.5B	MATH500 (500)	mean of 16 samples	−0.1 points
5.4	DeepSeek-Prover-V1.5	LeanWorkbook (89K)	cumulative, training set	+15.3 points
5.5	Qwen2.5-7B	HealthBench-1k (1,000)	greedy	+12.0 points
5.6	Qwen2.5-Math-7B	6 benchmarks	mean of 8–32 samples	+3.6 points
6.1a	Qwen2.5-Math-1.5B	MATH500 (500)	greedy, test-chosen step	180 → 368 problems
6.1b	Qwen2.5-7B	4 benchmarks	mean of 8–64 samples	+0.2 points
6.3a	s1-32B	AIME24 (30)	greedy	15 → 17 problems
6.3b	Qwen2.5-Math-7B	AIME24 (30)	mean of 16 samples	≈4 → 12 problems

Table B.2: **How large each difference in Table B.1 is.** Rows carry the keys of Table B.1. Problem counts convert scores into problems solved. Runs are repeated training runs; samples are repeated generations at evaluation. –: not applicable; not stated: not stated by the source.

Section 7 notes that in its three examples the questions stay the same. Two of the seven pairs share their questions. The random-reward pair comes from one study with the model as the sole change, the cleanest evidence that the learner matters; both sides of the unsolved-problem pair train on DAPO-Math-17K. The other five also differ in question pool, benchmark or budget, as the cards of Appendix B.3 show for the teacher pair, s1 and OpenThoughts; the teacher example of Section 7 is instead the nine teachers of Hu et al. (2026), compared on the same questions.

B.3 Two Filled Reporting Cards

A reporting card lists, for one result, the learner, the target and the update rule that Section 7 asks reports to state, together with the route, the checker, the budget and the source of the questions. Table B.4 fills in cards for the two conflicts that the experiments proposed in Section 7 target, and in both pairs more differs than the factor marked in Table B.3. The teacher studies also differ in budget and question pool; neither applies a checker, so s1, for example, keeps traces whether or not their answers are right (Table B.6). The studies of unsolved problems differ in learner as well as route. Complete systems leave several of these fields blank (Table D.2).

					Differs in	
§	**Topic**	**One side**		**Other side**	**Learner**	**Route**
4.1	Random rewards†	Help Qwen	vs.	fail on Llama, OLMo	●	
4.2	Process reward	Wins at reranking	vs.	gamed in RL		●
4.3	Unsolved problems⋆	Dropped in RL	vs.	teach most in distillation		●
4.4	Diversity	New tasks help	vs.	reworked pool: no gain	*intervention*	
5.2	Stronger teacher⋆	R1 beats Gemini	vs.	QwQ beats R1	●	
5.6	SFT with RL	Interleaved helps	vs.	stacked interferes		●
6.2	New capability	Base catches up at large k	vs.	long RL solves new tasks	●	

Sources, one side / other side: 4.1 Shao et al., 2026; 4.2 Lightman et al., 2023 / Cheng et al., 2025; 4.3 Yu et al., 2025 / Hou et al., 2026; 4.4 He et al., 2026 / Shin et al., 2026; 5.2 Muennighoff et al., 2025 / Guha et al., 2025; 5.6 Ma et al., 2025 / Zhu et al., 2026; 6.2 Yue et al., 2025 / Liu et al., 2025b.

Table B.3: **Seven conflicts and what separates them.** §: main-text section that raises the conflict. A dot marks the most salient difference; *intervention*: the two sides change the data in different ways, so neither learner nor route is marked. † Both sides come from one study. ⋆ Targeted by an experiment proposed in Section 7.

	Who writes the traces? §5.2		**Are unsolved problems useful?** §4.3	
	s1 Muennighoff et al., 2025	OpenThoughts Guha et al., 2025	DAPO Yu et al., 2025	OPD Hou et al., 2026
Learner	Qwen2.5-32B-Inst.	Qwen2.5-7B-Inst.	Qwen2.5-32B base	R1-Distill-1.5B
Target	Gemini or R1 traces	R1 or QwQ traces	answer reward	teacher's tokens
Route	SFT	SFT	RLVR	OPD
Update rule	likelihood	likelihood	weighted likelihood	per-token divergence
Checker	none	none	answer match	none
Budget	1K examples	31.6K examples	512 × 16 per step	64 × 8 per step
Questions	1K of 59K, 16 sources	OpenThoughts pool	DAPO-Math-17K	DAPO-Math-17K
Result	R1 beats Gemini	QwQ beats R1	unsolved dropped	hard ones teach most

Table B.4: **Two filled reporting cards.** Each column is one side of a conflict in Table B.3. §: main-text section that raises the conflict. Grey: the two sides agree. Budget: examples for SFT; prompts × samples per prompt for RL and OPD.

B.4 Random Checker Errors Weaken Training; Systematic Ones Redirect It

Random checker errors shrink the expected training update without changing its direction, whereas systematic errors change what is rewarded, which explains why random errors seem to average out in Section 7 while errors the learner can exploit do not. We call a checker error *random* if whether it occurs depends only on the response's correctness and is redrawn at every step, and *systematic* otherwise; in these terms, the errors the learner can exploit in Section 7 are systematic false accepts.

Write $r(y) \in \{0, 1\}$ for whether a response y is correct, $\tilde{r}(y)$ for the checker's reward, and α and β for the probabilities that the checker accepts a wrong response and rejects a correct one. Random errors satisfy two assumptions: (i) α and β depend on y only through $r(y)$, and (ii) the errors are drawn afresh at every step. Under both,

$$\mathbb{E}[\tilde{r}(y)] = \alpha + (1 - \alpha - \beta)\, r(y), \tag{2}$$

a shifted and shrunken copy of the true reward, so the expected update of a plain policy gradient points the same way as with a perfect checker, scaled by $1 - \alpha - \beta$ (0.8 at $\alpha = \beta = 0.1$); GRPO's normalization and clipping (Appendix C.1) make this only approximate.

Plesner et al. (2026) test this regime: they flip unit-test outcomes with equal false-accept and false-reject rates and redraw the flips each time a problem is revisited. The 15% cited in Section 7 comes from their sweep of group-level flips, which apply one draw to all responses to a prompt; these keep peak accuracy within 2 points of the clean run up to that rate. At a 10% rate, Qwen3-8B's best MBPP pass rate is 0.89 with group-level flips and 0.85–0.86 with per-sample flips (two variants of each, the four ways of Table B.5), against 0.90 without noise and 0.49 before training; a small LLM judge that often accepts wrong code does far more harm (Table B.5).

Coin-flip rewards, the random rewards of Section 4.1, are the limit $\alpha = \beta = \frac{1}{2}$, where the expected reward no longer depends on correctness, and Shao et al. (2026) confirm that without GRPO's clipping the expected update is zero. With clipping, updates favor what the model already does often: Qwen2.5-Math-7B writes its reasoning as unexecuted code in 65% of its MATH-500 answers before training and in over 90% after, and before training these answers are correct 60.9% of the time against 35.0% for the rest. The gain thus comes from the update rule meeting the learner's prior rather than from the labels, which is why it appears for Qwen and not for Llama or OLMo.

Systematic errors break assumption (i) or (ii). If a judge accepts any response with some feature f, such as consisting only of "Thought process:", then in the extreme case $\tilde{r}(y) = r(y) + (1 - r(y))\, \mathbf{1}[f(y)]$, and the policy can raise its reward by producing f whether or not it is right (Zhao et al., 2025b). A model's own majority vote breaks (i) differently: the label is the policy's most common answer, so its errors follow the policy's beliefs (Huang et al., 2026; Pan et al., 2026), although early in training most rewards can stay correct because wrong answers rarely agree with one another (Zuo et al., 2025). Assumption (ii) fails for a wrong reference answer fixed to a question, which is systematic even if the corrupted questions were chosen at random.

False rejects are systematic but not exploited. A checker that rejects correct answers with a feature g, such as $x = 2$ written as $4/2$ (attempt a_4 in Table C.1), gives $\tilde{r}(y) = r(y)\,(1 - \mathbf{1}[g(y)])$; the policy cannot gain reward through g, but its correct answers are pushed away from that format, a loss that grows as the policy gets more answers right (Huang et al., 2025). Apart from coin-flip rewards, every harmful error in Table B.5 is systematic. No study we found injects random and systematic errors into one checker at matched rates, the test proposed in Section 7; Plesner et al. (2026) themselves call for varying false-accept and false-reject rates independently, including errors that are not redrawn.

B.5 Correct Traces Can Teach Less, and Wrong Ones Can Teach

Section 5.2 concludes that a good teacher writes traces the student can learn from, not merely traces that are correct; Table B.6 splits the evidence behind that conclusion into studies that fix the writer and vary correctness, and studies that keep every trace correct and vary the writer. Correctness proves less necessary than expected in the first group, where traces with wrong final answers score only 3.2 points below correct ones (Li et al., 2025), and insufficient in the second, where correct traces written with the solution in view cut accuracy by more than half (Kim et al., 2026). STaR is the counterexample: its traces are written with the answer given as a hint and kept only if correct, yet they raise GPT-J's accuracy (Zelikman et al., 2022). STaR and Kim et al. (2026) differ in what the

§	Checker	Error	Effect
RANDOM			
7	Unit tests, random flips Plesner et al., 2026	10% flip rate, four ways	✓ 0.85–0.89 vs. 0.90 clean
RANDOM, COIN-FLIP LIMIT			
4.1	Coin flip Shao et al., 2026	ignores the answer	✓ Qwen ✗ Llama, OLMo
SYSTEMATIC FALSE ACCEPTS			
7	Small LLM judge Plesner et al., 2026	often accepts wrong code	✗ 0.70 vs. 0.90 clean
4.1	LLM judges Zhao et al., 2025b	":" accepted up to 94%	✗ answers "Thought process:"
4.1	Fine-tuned LLM verifier Huang et al., 2025	accepts garbled patterns	✗ hacked during RL
4.2	Neural PRM DeepSeek-AI et al., 2025	not stated	✗ hacked in large-scale RL
4.2	Summed step rewards Cheng et al., 2025	not stated	✗ training collapses
3	Visible unit tests Zhao et al., 2026	pass without the goal	✗ 97% visible, 0% held-out
SYSTEMATIC, OWN MAJORITY VOTE			
6.3	Test-time majority vote Zuo et al., 2025	label error 63%	✓ rewards still 92% correct
5.4	Self-play majority vote Huang et al., 2026	label error 21% → 37%	✗ gains, then decline
5.4	Self-consistency reward Pan et al., 2026	not stated	✗ reinforces own errors
SYSTEMATIC FALSE REJECTS			
4.1	Rule-based verifiers Huang et al., 2025	14% of correct answers	✗ harm grows as policy improves
4.1	Rule-based verifier Xu et al., 2025c	over 38% of rejections	✗ fixing them helps RL
SYSTEMATIC, BOTH DIRECTIONS			
5.5	One score per response Xie et al., 2026	ignores the steps	✗ 18% wrong steps rewarded

Table B.5: **Which checker errors do harm.** Group headers name each kind of error; the text above derives its effect on the expected update. ✓: training still improves; ✗: training is harmed or hacked. §: main-text section that uses the row; not stated: the source gives no error rate. Rows differ in model, task and metric, so rates are not comparable across rows.

writer sees (the answer or a full solution), model, trace length and task, and no study we found tests hinted traces in both settings.

Study	Learner	Result
CORRECTNESS VARIED, WRITER FIXED		
Wrong final answers Li et al., 2025	Qwen2.5-32B, 4.6K traces	63.1 vs. 66.3 correct; base 56.7
Wrong, reflective primers Gandhi et al., 2025	Llama-3.2-3B, Countdown	RL gains match correct primers
Unchecked answers Muennighoff et al., 2025	Qwen2.5-32B	54% (s1K), 63% (s1K-1.1)
Answer filtering Guha et al., 2025	Qwen2.5-7B	40.0 LLM-verified vs. 41.6 random
WRITER VARIED, EVERY TRACE CORRECT		
Solution in view Kim et al., 2026	R1-Distill-7B, 800 traces	20.2 vs. 51.0 unguided; base 54.8
Nine teachers Hu et al., 2026	Qwen2.5 0.5B, 3B, 7B	stronger teacher, not better student
QwQ vs. R1 teacher Guha et al., 2025	Qwen2.5-7B	71.6 vs. 64.8
Answer as a hint Zelikman et al., 2022	GPT-J	72.5 vs. 68.8 without hinted traces
Own vs. human solutions Singh et al., 2024	PaLM 2	own filtered solutions train better

Table B.6: **Correct traces can teach less, and wrong ones can teach.** Metrics: average over four math benchmarks (wrong final answers), eight math, code and science benchmarks (answer filtering) or three math benchmarks (QwQ vs. R1); Countdown score (primers); share of traces the authors' grader judges correct (unchecked answers); AIME 2024 (solution in view); accuracy on DeepScaleR, GSM8K and MATH (nine teachers); CommonsenseQA (hint); MATH and HumanEval (own vs. human). The filtering row compares two equal-size subsets of 63.2K answers, one accepted by a GPT-4o-mini judge and one drawn at random; training on all 63.2K scores 41.9, a different configuration from the 41.7 baseline in Table B.1. Qwen2.5 learners are Instruct models.

B.6 SCALING RESULTS COUNT DIFFERENT RESOURCES

The few-prompt results of Section 6.1 still draw many samples and pay costs that the prompt count hides (Table B.7). One-shot RLVR copies its single prompt to fill a batch of 128 and samples 8 responses to each copy: 1,024 attempts per step, about two million over 2,000 steps (Wang et al., 2025). Choosing the prompt took a 500-step run, and the authors note that the method does not save RL compute. A reward for answer format alone reaches 65.0 on MATH500 (65.6 at that run's best MATH500 checkpoint), against 36.0 before training and 73.6 after one-shot RLVR. The 65.0 and 73.6 come from the checkpoint with the best average on the six test benchmarks. Under OPD, eight hard prompts reach 53.6 against 53.7 for the full 17K-prompt set (Hou et al., 2026). A one-prompt run keeps the full-set batch of 64 prompts with 8 responses each for the full set's 279 steps (one epoch), so its prompt fills all 512 samples of every step; eight prompts share them. Reuse has limits: repeating the data up to 25 times causes no significant loss, whereas 100 times overfits (Tan et al., 2026). Khatri et al. (2025) find that many design choices, such as loss aggregation and curriculum, mainly change how fast a scaling curve rises, not the level it approaches, so a scaling claim should say which of the two moved.

B.7 FINDING VERSUS SELECTING, AND WHY RL CAPABILITY STUDIES DISAGREE

Section 6.2 leaves open whether RL expands what a model can solve; this subsection derives its decomposition and shows that the studies behind the debate count different events (Table B.8). The decomposition $P(D_k) = P(A_k)\,P(D_k \mid A_k)$ holds because a selected answer can only be correct if some candidate is, so D_k lies inside A_k. For a problem solved with probability p per attempt, $P(A_k) = 1 - (1-p)^k$; from $n \geq k$ samples with c correct, its unbiased estimate is $1 - \binom{n-c}{k} / \binom{n}{k}$ (Yue et al., 2025). The selector sets $P(D_k \mid A_k)$: a perfect checker delivers all of $P(A_k)$, whereas a random pick delivers only p whatever k is (with $p = 1/64$, 0.016 against $P(A_{64}) \approx 0.64$); p is also what mean accuracy over k samples estimates. Majority vote (the most common answer) and best-of-N (a scorer's top answer) also count D_k. Majority vote can lock onto a wrong answer: if a learner answers $2x + 3 = 7$ with $x = 2$ on 30% of attempts and with $x = 5$ on 50%, $P(A_k)$

Study	Prompts	Per step	Steps	Hidden cost
RL AND ON-POLICY DISTILLATION				
One-shot RLVR Wang et al., 2025	1	128×8	2,000	500-step run to rank 1,209 problems
One-prompt OPD Hou et al., 2026	1	64×8	279	2×16 samples on 1,000 problems
DAPO Yu et al., 2025	17K	512×16	not stated	over-sampling to fill batches
AR3PO Zhang et al., 2025a	14K	512×5.7	not stated	reuses earlier correct responses
SUPERVISED FINE-TUNING				
s1 Muennighoff et al., 2025	1K	16×1	315	59K traces; 2 models graded on 52K
OpenThoughts3 Guha et al., 2025	≈75K	not stated	not stated	22K H100-hours to annotate

Table B.7: **What each scaling result counted.** Per step: prompt slots in the batch $\times$ samples per slot; a one-prompt run fills every slot with copies of its prompt, and an SFT sample is a trace. OpenThoughts3 states neither factor nor a step count: it trains for 5 epochs on batches of 512 packed sequences, each of which can hold several traces. OPD ranks problems by difficulty with 16 samples each from the student and the teacher. AR3PO's 5.7 is an average, because it adapts the number of samples to each prompt; its 4.2-fold saving is measured against its own DAPO run, which drew $1{,}536 \times 8$ per step. not stated: not stated by the source.

approaches 1 as k grows while the majority answer approaches 5. This is also how majority-vote labels can reinforce errors (Appendix B.4).

Metric	Learner	Event counted	Finding
pass@k Yue et al., 2025	public RL models	A_k, $k \leq 1024$	base overtakes RL at large k
Support Wu et al., 2026	ProRL-1.5B	probability $> \epsilon$	math: 5 gained, 56 lost
pass@128 Liu et al., 2025b	ProRL-1.5B, 2K+ steps	A_{128}	gains largest where base is weak
CoT-pass@k Wen et al., 2025	DAPO-Qwen-32B	A_k and reasoning	ahead at all $k \leq 1024$ on AIME
pass@(k, T) Zhai et al., 2026	Qwen2.5-7B agent	A_k within T turns	expands on compositional tasks

Table B.8: **Studies of whether RL expands what a model can solve.** Each row compares an RL-trained learner with the model it started from. A_k: at least one of k candidates is correct; support: the correct answers whose probability exceeds a small cutoff ϵ; CoT-pass@k counts A_k only if a judge also accepts the reasoning. Rows differ in base model, training length, metric and task, so findings are not directly comparable.

Averaged over problems, pass@1 is the mean per-problem p, whereas pass@k at large k approaches the share of problems with $p > 0$, so the two can move in opposite directions. For example, training that moves two problems from $p = (0.30,\ 0.02)$ to $(0.90,\ 0)$ raises pass@1 from 0.16 to 0.45 but lowers pass@64 from 0.86 to 0.50, the pattern Yue et al. (2025) report for RL. Finite budgets blur the line: a problem solved once in 1,000 attempts is missed by 256 samples 77% of the time, so an answer that was not sampled does not show that the model cannot produce it.

Two of the opposing studies examine the same model and partly agree: Wu et al. (2026) find that ProRL-1.5B, the checkpoint of Liu et al. (2025b), gains 5 reachable answers on math and loses 56, and ProRL itself reports minimal or negative pass@128 gains where the starting model already does well, as on some math benchmarks. Distillation from a stronger teacher, by contrast, lifts the whole pass@k curve above the base model (Yue et al., 2025). Because the rows of Table B.8 differ in the event counted as well as in learner, training length and task, they do not settle the question.

C WORKED EXAMPLES

Appendix B reports what studies found, with the derivations of Appendices B.4 and B.7; this section works through three further mechanisms of Sections 2 and 4 on constructed examples. The main example is the equation $2x + 3 = 7$ of Section 4.2, with the four attempts in Table C.1; Appendix C.3

adds constructed problems with a given solve rate. Appendix C.1 supports Section 2 and the route column of Table B.3, Appendix C.2 supports Section 4.2, and Appendix C.3 supports Section 4.3.

	Attempt	**Answer**	**What is wrong**
a_1	$7-3=4$, so $2x=4$, so $x=2$	2	nothing
a_2	$7+3=10$, so $2x=10$, so $x=5$	5	the first step
a_3	$7-3=4$, so $2x=5$, so $x=2$	2	a wrong step, then the right answer
a_4	$2x=4$, so $x=4/2$	$4/2$	nothing, but a strict string match rejects it

Table C.1: **Four attempts at "Solve $2x+3=7$".** The attempts are constructed for illustration; a_3 is the example of Section 4.2.

C.1 One Record under Different Routes

What a record teaches belongs to the learner and route, not to the record alone (Section 2). The four attempts of Table C.1 show how the route changes it, which is the difference the route column of Table B.3 marks. A *route* is the recipe by which a record reaches the learner, such as SFT, rejection sampling, RLVR, on-policy distillation (OPD) or selection only (best-of-N). Choosing a route fixes the *target*, what the loss pulls the learner toward, and the update rule. We use $2x+3=7$ rather than the Game of 24 record of Figure 1a because its attempts include a right answer after a wrong step (a_3) and a correct answer that a strict checker rejects (a_4). Table C.2 adds to the routes of Figure 1b the majority-vote reward of Section 5.4 (Huang et al., 2026; Pan et al., 2026) and the rubric reward of Section 5.5 (Gunjal et al., 2025).

Under GRPO, each attempt is weighted by its reward minus the group mean, divided by the standard deviation of the group's rewards (Shao et al., 2024). A strict string match gives rewards $(1,0,1,0)$ and weights proportional to $(+1,-1,+1,-1)$, so the wrong step in a_3 is reinforced as much as a_1, and the correct a_4 is pushed down as hard as the wrong a_2. A checker that accepts "$4/2$" gives rewards $(1,0,1,1)$ and weights proportional to $(+1,-3,+1,+1)$. If all four attempts were correct, or all wrong, every weight would be zero; this is why DAPO drops such groups (Section 4.3). Rejection sampling with the strict match keeps a_1 and a_3, and so imitates the wrong step. OPD instead scores every token of every attempt against the teacher's next-token distribution, so it can correct a_2 at its first step even when no attempt in the group is right (Agarwal et al., 2024; Hou et al., 2026).

Route	**Learner**	**Target**	**Update rule**	**A wrong attempt**	$p=0$
SFT	student	teacher trace	likelihood	unused	✓
Rejection sampling	student	own passing attempts	likelihood	discarded	✗
RLVR	policy	higher reward	weighted likelihood	pushed down	✗
OPD	student	teacher's tokens	per-token divergence	corrected per token	✓
PRM training	PRM	step labels	classification	negative labels	–
Best-of-N	–	–	selection only	not chosen	–
Majority vote	policy	majority answer	weighted likelihood	rewarded if common	✗
Rubric reward	policy	criteria met	weighted likelihood	partial credit	✓

Table C.2: **One record, eight routes.** The first six are the routes of Figure 1b. p: the learner's per-attempt success rate on a problem; column $p=0$: whether a problem the learner never solves still gives a useful signal (✓ yes, ✗ no); –: not applicable.

C.2 What a Step Label Measures

The three process labels of Section 4.2 ask different questions about a step. Two of them depend on a second model: the Math-Shepherd label on the model that samples completions from the step (the completer), and the progress label on the model whose success rate it measures (the prover). Table C.3 lists the three label types and three further questions a process score answers.

Take the second step of a_3, which writes $2x = 5$ after $7 - 3 = 4$. A PRM800K annotator would mark it bad (Lightman et al., 2023). Math-Shepherd labels a step good if any completion sampled from it reaches the reference answer, or scores it by the share that do (Wang et al., 2024); a completer that recomputes from $7 - 3 = 4$ reaches $x = 2$, so the step can count as good. A progress label, the change in a prover's success rate across the step, is negative if the prover succeeds less often after $2x = 5$ (Setlur et al., 2024). The same step is thus bad under the human label, good under the Math-Shepherd label, and bad under the progress label if the prover's success rate drops. A prover that succeeds from any step gives every step a progress near zero, and one of intermediate strength works best (Setlur et al., 2024).

Trained PRMs inherit the disagreement. On candidates from the same policy, PRMs trained on four label sets reach best-of-8 accuracies of 64.3 to 65.9, yet their ProcessBench F1 at locating the first error ranges from 28.9 with Math-Shepherd's labels to 56.5 with human ones, although the human labels cover fewer, different responses (Zhang et al., 2025b). How step scores are combined matters as well (Section 4.2): step scores of $(0.9, 0.2, 0.9)$ for a_3 sum to 2.0 with a minimum of 0.2, and two added thinking steps at 0.95 raise the sum to 3.9 but leave the minimum at 0.2. Cheng et al. (2025) observe this failure: under summed rewards, the policy learns to output only thinking steps.

Question	**Label or measure**	**Depends on**
LABEL TYPES (SECTION 4.2)		
Is this step valid? Lightman et al., 2023	human rating: good, neutral or bad	annotators
Can this prefix still succeed? Wang et al., 2024	any or share of completions right	completer, samples
Did this step help? Setlur et al., 2024	change in a prover's success rate	prover
OTHER QUESTIONS A PROCESS SCORE ANSWERS		
How are step scores combined? Lightman et al., 2023; Cheng et al., 2025	product to rerank; sum or minimum in RL	use of the score
Will this score select well? Zhang et al., 2025b	best-of-N accuracy	candidate pool
Is the trace faithful? Turpin et al., 2023; Lanham et al., 2023	effect of edits to the input or the trace	task, model

Table C.3: **Six questions a process score can answer.** Two of the three label types depend on a second model: the completer or the prover.

C.3 WHICH PROBLEMS CARRY SIGNAL

Under RLVR, a problem carries signal only if the learner solves it on some attempts but not all (Section 4.3); how many problems do so depends on the sampling budget as well as on the learner. If the learner solves a problem with probability p per attempt, the probability that all G attempts in a group receive the same reward, and so give no signal under GRPO (Appendix C.1), is $p^G + (1-p)^G$. Table C.4 shows how quickly the share of groups with no signal grows as p moves away from 0.5: with 16 attempts, a problem solved 95% of the time gives no signal in 44% of groups.

	Per-attempt success rate p (or $1 - p$)			
Attempts per group	0.5	0.9	0.95	0.99
$G = 8$	0.8%	43%	66%	92%
$G = 16$	<0.01%	19%	44%	85%

Table C.4: **Share of groups that give no signal.** Computed as $p^G + (1-p)^G$; G: attempts per group; p: per-attempt success rate.

For the same reason, the share of problems reported as solved on every sample, or on none, depends on the number of samples. Skywork-OR1 draws 16 per problem (Figure 2c; He et al., 2025a); with 8, a problem solved 90% of the time would look solved on every sample in 43% of groups rather than 19%. This is why Table 1 asks reports to state how pass rates were estimated. The route matters too

(column $p = 0$ of Table C.2): rejection sampling needs one of its k attempts to pass, which happens with probability $1 - (1 - p)^k$, whereas OPD gains even on problems that neither student nor teacher solves (Hou et al., 2026).

D Reference Material

This section collects reference material for the main text: a glossary (Appendix D.1), the public releases of Section 3 (Appendix D.2), and the complete training systems whose pipelines Section 5 omits (Appendix D.3).

D.1 Glossary

Table D.1 lists the terms used in the main text and the appendix in alphabetical order, each with a short definition and the section that defines it or, for three standard terms the text never defines, first uses it.

D.2 Public Releases by Record Form and Check

Figure D.1 supports the point of Section 3 that what a record retains does not fix how its feedback is checked. It places twelve public releases on both dimensions of that section (Albalak et al., 2025; Yu et al., 2025; Yuan et al., 2025a; Arora et al., 2025; He et al., 2025b; Xu et al., 2025d; Moshkov et al., 2025; Guha et al., 2025; Wang et al., 2024; Lightman et al., 2023; Pan et al., 2025; Trivedi et al., 2024). Among releases with full solutions, DeepMath-103K and KodCode are checked programmatically, OpenMathReasoning combines a programmatic check with an LLM judge, and OpenThoughts3 keeps its traces without any check.

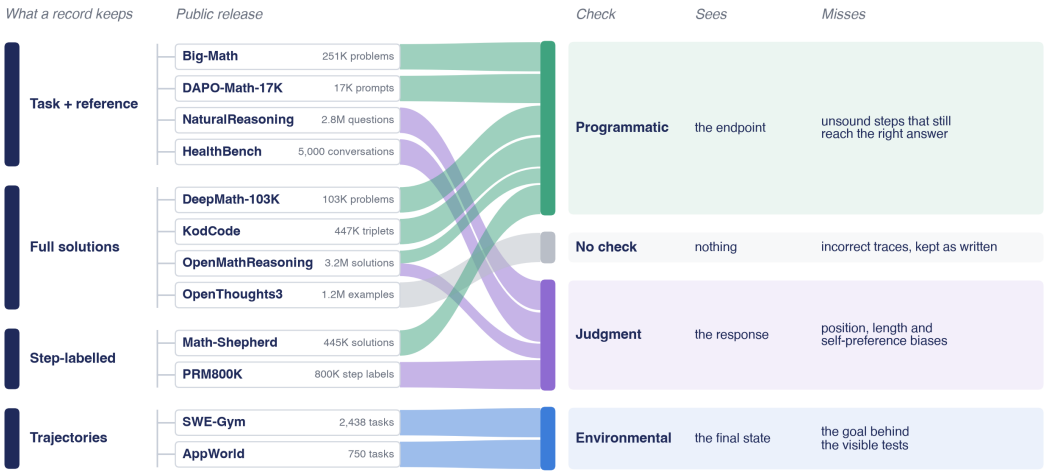

Figure D.1: **Public releases by what a record keeps and how its feedback is checked.** Each band links a release to the check it relies on; OpenMathReasoning uses two. Scales are as reported by each source. *Sees*: what the check inspects; *Misses*: errors that can pass it. The biases listed for judgment apply to LLM judges; PRM800K's step labels come from human annotators.

D.3 Complete Training Systems

Section 5 follows construction choices rather than systems, because a system name does not say which variable moved. Table D.2 gives the complete pipelines omitted from that discussion, grouped by how traces and rewards enter post-training. Rewards for reasoning are mostly rule-based checks; learned reward models appear for math in DeepSeekMath and Kimi k1.5, and for general tasks in the last stages of DeepSeek-R1 and Qwen3. Kimi k1.5 and MiniMax-M1 do not say where their SFT traces come from, and Qwen3-Coder reports no SFT stage.

Term	Meaning	§
Base model†	a pretrained model before post-training	1
Best-of-N selection	a scorer picks one of N answers; no model is updated	2
Checker	a program, environment or judge that decides if an attempt is correct	3
Environmental check	inspects the state an interaction leaves behind	3
False accept, false reject	a checker passes a wrong answer, or fails a right one	4.1
Feedback	a label or score attached to a record	2
Finding (A_k), pass@k	at least one of k attempts is correct	6.2
GRPO	RL that weighs each attempt against its group's mean reward	C.1
Informative problem	one the learner solves on some attempts but not all	4.3
Intervention	what a study does to the data, e.g. adding or reweighting tasks	B.2
Judgment-based check	scores a response against rubrics or preferences	3
Learner	the model that is updated: the student in distillation, the policy in RL	2
Learning interface	the learner, its training target and its update rule	2
Likert judging†	a judge rates a response on a scale, without a rubric	5.5
Loss masking	keeping tokens in the input but out of the loss	5.3
Majority vote, pseudo-label	the most common sampled answer, used as a label or as the final answer	5.4
Monte-Carlo step label	whether or how often completions from a step reach the answer	3
Off-policy	trained on attempts made by another model	2
On-policy	trained on the learner's own attempts	2
On-policy distillation (OPD)	the student samples; a teacher scores each of its tokens	2
Pass rate	the share of attempts that solve a problem	4.3
Process reward model (PRM)	a model that scores each step of a solution	2
Programmatic check	compares answers, runs tests or verifies proofs	3
Provenance	who produced each part of a record	2
Prover	a model whose success rate defines a progress label	C.2
Random, systematic error	a checker error is random if its occurrence depends only on the response's correctness and is redrawn every step; otherwise it is systematic	B.4
Record	a question, trace and answer, plus optional feedback and provenance	2
Rejection sampling	imitating the learner's own attempts that pass a checker	2
Reporting card	a result's learner, target, update rule, checker, budget and questions	B.3
Reward hacking†	raising the reward without solving the task	4.2
RLVR	RL with verifiable rewards, from an automatic check of the answer	2
Route	the recipe by which a record reaches the learner, such as SFT, rejection sampling, RLVR, OPD or selection only; it fixes the target and update rule	C.1
Rubric	a list of criteria that a good response should meet	5.5
Selection (D_k)	the answer picked from k attempts is correct	6.2
Self-play	a model writes its own training tasks	5.4
Target	what the loss pulls the learner toward	C.1
Teacher, student	in distillation, the model that supplies the target, and the learner	2
Trace	the intermediate reasoning steps that lead to an answer	2

Table D.1: **Glossary.** §: section that defines the term, in the main text or the appendix. † A standard term that the text uses without defining; § gives its first use.

System	Stages	Traces from	Reward
SFT, THEN RL			
DeepSeek-R1 DeepSeek-AI et al., 2025	(SFT → RL) × 2	R1-Zero, own RL	rules; RMs in last stage
DeepSeekMath Shao et al., 2024	SFT → RL	math solutions	learned RM
Phi-4-reasoning-plus Abdin et al., 2025	SFT → RL	o3-mini	answer check
Kimi k1.5 Team et al., 2025	SFT → RL	not stated	chain-of-thought RM, tests
Qwen3, large Yang et al., 2025	(SFT → RL) × 2	QwQ-32B, own RL	verifiers, judge, RM
Magistral Small Mistral-AI et al., 2025	SFT → RL	Magistral Medium	rule checks
RL WITHOUT TRACE SFT			
R1-Zero DeepSeek-AI et al., 2025	RL on base	–	rule checks
Open-Reasoner-Zero Hu et al., 2025	RL on base	–	answer match
DAPO Yu et al., 2025	RL on base	–	answer match
Magistral Medium Mistral-AI et al., 2025	RL on instruct	–	rule checks
DISTILLATION ONLY			
Qwen3, small Yang et al., 2025	distillation	large Qwen3	teacher's outputs
AGENT ENVIRONMENTS			
MiniMax-M1 MiniMax et al., 2025	SFT → RL	not stated	rules, sandboxes, judge
Qwen3-Coder Qwen Team, 2025	code RL, agent RL	–	execution feedback

Table D.2: **Complete training systems.** Listed for orientation; no controlled comparison in Section 5 rests on them, because each differs from the others in more than its optimizer. × 2: two rounds; RM: learned reward model; not stated: the report does not say; –: no SFT stage reported.