LEARNING FROM SYNTHETIC DATA IMPROVES MULTI-HOP REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning (RL) has been shown to significantly boost reasoning capabilities of large language models (LLMs) in math, coding, and multi-hop reasoning tasks. However, RL fine-tuning requires abundant high-quality verifiable data, often obtained through human-annotated datasets and LLM-as-verifier loops. Both of these data types have considerable limitations: human-annotated datasets are small and expensive to curate, while LLM verifiers have high scoring latency and are costly to operate. In this work, we investigate the use of synthetic datasets in RL fine-tuning for multi-hop reasoning tasks. We discover that LLMs fine-tuned on synthetic data perform significantly better on popular real-world question-answering benchmarks, even though the synthetic data only contain fictional knowledge. On stratifying model performance by question difficulty, we find that synthetic data teaches LLMs to *compose knowledge*, which we to be a fundamental and generalizable reasoning skill. Our work thus highlights the utility of synthetic reasoning datasets in improving LLM reasoning capabilities.

1 INTRODUCTION

Reinforcement learning (RL) has demonstrated remarkable success in enhancing the reasoning capabilities of large language models (LLMs) across diverse domains including mathematics, coding, and complex reasoning tasks (Bai et al., 2022; Shao et al., 2024; Lambert et al., 2025; Guo et al., 2025a; Guan et al., 2025). Answering questions in these language model reasoning domains characteristically require executing *multi-hop* or *multi-step* solution trajectories to reach the final solution. That is, the questions require solving intermediate subproblems in math and coding tasks, or sequencing deduction steps in natural language question-answering. In recent years, most LLM-oriented reasoning benchmarks reflect this multi-hop structure (Mirzadeh et al., 2025; MAA; Yang et al., 2018; Trivedi et al., 2022) and are popularly used to evaluate gains in LLM reasoning capabilities. Beyond evaluation, these datasets have also been used to fine-tune LLMs (Shao et al., 2024; Rafailov et al., 2023), demonstrating that they are valuable resources for boosting LLM reasoning capabilities.

However, the effectiveness of RL fine-tuning for LLM reasoning is fundamentally constrained by the availability of high-quality training data of questions and verifiably correct answers (Lambert et al., 2025). Curation of new datasets is both time-consuming and expensive, especially when reasoning tasks require reliable ground-truth labels (Xie et al., 2024). In addition, as LLMs are trained at internet-scale, they eventually become prone to data leakage, memorization, and thus reasoning improvements are often unreliable (Gong et al., 2025; Xie et al., 2024). As an observed consequence, the pace of language model training is starting to outpace the availability of high-quality human-written text required for reasoning training (Villalobos et al., 2024; Muennighoff et al., 2023). In response, it has been a recent practice to leverage *synthetic data* in both LLM pretraining and fine-tuning, either by expanding upon the existing data with generated traces (Trinh et al., 2024; Ruan et al., 2025) or by training on synthetic problems generated from stronger model (Abdin et al., 2025). Another major trend focuses on generating problems within systematically verifiable domains such as mathematics and coding, enabling the application of reinforcement learning with verifiable reward signals (RLVR; Guo et al., 2025a; Lambert et al., 2025). While this approach has shown promising results, reasoning domains beyond mathematics and coding remain relatively underexplored, primarily due to the inherent challenges in establishing verifiable systems in more general problem settings (Su et al., 2025).

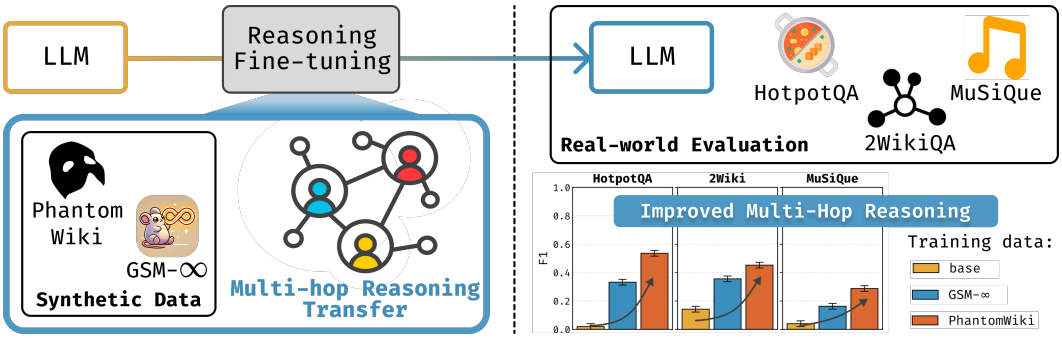

Figure 1: We analyze performance transfer from synthetic to real-world multi-hop reasoning.

In this work, we investigate a fundamental research question: ***can models develop general reasoning capabilities solely from synthetic data, without relying on real-world knowledge?*** Among various reasoning capabilities, we focus specifically on *knowledge composition*—the fundamental ability to integrate information across multiple inferential steps for multi-hop reasoning.

To answer this question, we examine whether capabilities acquired from synthetic multi-hop reasoning datasets can effectively generalize to real-world natural language question-answering scenarios. Specifically, we conduct systematic experiments using knowledge composition datasets such as PhantomWiki (Gong et al., 2025) and GSM-Infinite (GSM-∞) (Zhou et al., 2025), to fine-tune LLMs with reinforcement learning across diverse reasoning complexity levels. Our findings demonstrate that these synthetic datasets provide scalable and verifiable training signals, enabling successful transfer of enhanced reasoning capabilities to real-world question-answering benchmarks. Training on these synthetic datasets consistently improves performance on in-context multi-hop reasoning tasks such as HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022). Moreover, performance transfer trends are consistent across model families and sizes, with no evidence of overfitting at scale. For example, Qwen3-0.6B model trained on PhantomWiki achieves F1 improvements of 62% on HotpotQA, 63% on 2WikiMultihopQA, and 132% on MuSiQue, relative to the base model. Scaling synthetic training data does not cause overfitting, demonstrating robust generalization. We further analyze model performance during training, stratified by question difficulty levels, which is systematically defined by the number of required reasoning steps. We establish that improvements on more challenging questions in synthetic datasets consistently translate to enhanced performance on more difficult real-world question-answering tasks, empirically demonstrating the transferability of multi-hop reasoning capabilities.

Our key insight is that reasoning capabilities developed on synthetic data—particularly the *ability to compose and chain logical inferences*—can generalize to real-world multi-hop reasoning scenarios, even when the training and evaluation domains share no factual overlap. Our contributions:

1. We propose synthetic multi-hop datasets as a scalable, cost-effective source of reasoning training data that provides infinite, verifiable training signals, demonstrating that multi-hop reasoning capabilities can be effectively learned from synthetic data without factual overlap between training and evaluation domains.
2. We provide empirical evidence for synthetic reasoning training that generalizing to real-world scenarios, demonstrating performance gains across model families and sizes, and establishing the practical viability of synthetic data for reasoning enhancement.
3. We study reasoning transfer to synthetic and real-world tasks across question difficulty levels, demonstrating that improvements on synthetic tasks with varying reasoning complexity translate to enhanced performance on increasingly challenging real-world tasks.

2 Background and Related Work

Reasoning in Large Language Models. While LLM reasoning is a long-standing research area, the definition and assessment of reasoning capabilities is ambiguous and therefore complex (Xie et al., 2024; Han et al., 2025). Thinking and reasoning models like DeepSeekMath (Shao et al., 2024),

DeepSeek-R1 (Guo et al., 2025a), or Phi-4-reasoning (Abdin et al., 2025) are typically evaluated in their reasoning skills through performance on various benchmarks. These benchmark tasks may range from technical and abstract domains like mathematics, algorithms, coding, puzzle-solving (Hendrycks et al., 2021; MAA; Cobbe et al., 2021; Jain et al., 2025; Chollet et al., 2025), to more knowledge-intensive domains like the sciences and law (Rein et al., 2024; Sawada et al., 2023), general common sense, abductive, and counterfactual reasoning (Talmor et al., 2019; Zhao et al., 2023; Bhagavatula et al., 2020; Wu et al., 2025a; Hüyük et al., 2025), natural language question-answering (Trivedi et al., 2022; Yang et al., 2018; Ho et al., 2020; Tang & Yang, 2024; Qi et al., 2021), and interaction with the environment through planning and tool use (Patil et al., 2024; Zhuang et al., 2023; Yao et al., 2024). Many of these benchmarks require breaking the question down into intermediate subproblems and composing them together to arrive at the correct final answer (Gong et al., 2025; Xie et al., 2025); this behavior is considered to be one of the intrinsic properties of effective reasoning models (Gandhi et al., 2025).

Training and Fine-tuning Large Reasoning Models. LLM performance and generalization on reasoning benchmarks can be improved with training or fine-tuning using several classes of techniques. The simplest approach is to train on the datasets directly using supervised fine-tuning (SFT; Lambert et al., 2025) with the next-token prediction objective. This includes variations to add more helpful instructions or to encourage a more detailed thinking process, for instance through instruction fine-tuning (Chung et al., 2024) and chain-of-thought (CoT) modeling (Xiang et al., 2025; Zelikman et al., 2022; Hao et al., 2025; Yao et al., 2023; Chen et al., 2023; Wan et al., 2025). reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ouyang et al., 2022) emerged as a more complicated, RL-based framework for fine-tuning models using human preferences. RLHF algorithms include policy gradient-based PPO (Schulman et al., 2017), and variants or simplifications like GRPO (Shao et al., 2024) and DPO (Rafailov et al., 2023), among others (Hu et al., 2025; Pang et al., 2024; Brantley et al., 2025; Yu et al., 2025; Liu et al., 2025; Shrivastava et al., 2025).

Many LLM reasoning benchmarks benefit from having objective ground-truth answers (such as the correct answer to a math question); replacing the reward model in RLHF with a procedural verification function has been termed reinforcement learning with verifiable rewards (RLVR; Lambert et al., 2025). While this technique has been utilized in several recent reasoning models (Lambert et al., 2025; Guo et al., 2025a; Abdin et al., 2025), its ability and mechanisms for eliciting fundamentally novel reasoning patterns remains an open research area (Wen et al., 2025; Yue et al., 2025; Shao et al., 2025; Zhao et al., 2025).

Leveraging Synthetic Data. Fine-tuning large reasoning models has several challenges. One challenge is that the abstract (multi-hop) reasoning skills may be difficult to isolate in any particular benchmark: they could be confounded both by other skills (such as arithmetic or writing syntactically-correct code) or memorization—in the way that allows the model to leverage required implicit knowledge while preventing it from recalling the memorized answer itself (Wu et al., 2025b; Xie et al., 2024; Yu et al., 2024). Moreover, as LLMs are trained at internet-scale, reasoning benchmarks gradually become prone to test set leakage (Gong et al., 2025; Wu et al., 2025b), while novel and unseen benchmarks with reliable rewards become more scarce. All these challenges can be alleviated using synthetic datasets, which can isolate specific reasoning aspects while providing potentially unlimited number of new examples with verifiable rewards.

Most synthetic reasoning benchmarks are generated programmatically, especially in mathematics (Mirzadeh et al., 2025; Zhou et al., 2025; Wu et al., 2025b), logic puzzles (Xie et al., 2024; Shojaee et al., 2025; Stojanovski et al., 2025), and some forms of natural language question-answering (Gong et al., 2025; Guo et al., 2025b; Sinha et al., 2019). Other benchmarks leverage LLMs to create additional examples and reasoning traces, augmenting existing curated datasets (Yang et al., 2025; Goldie et al., 2025; Huang et al., 2025; Saad-Falcon et al., 2024; Li et al., 2025).

However, as with RLVR, the effectiveness and applicability of these synthetic data to real-world reasoning skills remains an underexplored question (Yu et al., 2024; Mizrahi et al., 2025; Abbe et al., 2024b;a; Stojanovski et al., 2025), which we study in this work.

3 METHODOLOGY

To comprehensively study the transfer performance of synthetic to real-world datasets, we RL fine-tune LLMs of sizes varying from $\approx$ 0.75B to 4B parameters: Qwen3-0.6B, Qwen3-1.7B (Qwen Team, 2025), Qwen2.5-1.5B-Instruct (Qwen Team, 2024), and Phi-4-mini-reasoning (Abdin et al., 2025). We fine-tune each model for 1 epoch on a random shuffle of each of our synthetic training datasets, which required 4 NVIDIA H100 GPUs. We describe our synthetic training dataset selection, RL fine-tuning algorithm, prompt configuration and reward models below; see Appendix A for full implementation details.

3.1 SYNTHETIC TRAINING DATASETS

To fine-tune LLMs with reinforcement learning, recent works highlight the need for large datasets with two important characteristics: *scalable verification of model generations*, and *questions of varying difficulty* (Guo et al., 2025a; Wen et al., 2025; Shao et al., 2025; Lambert et al., 2025; Abdin et al., 2025). Scalable verification is essential for quickly determining the on-policy generation rewards, enabling reasonable fine-tuning times, meanwhile a mix of easy and hard questions helps provide both informative and meaningful training signal. With these criteria in mind, and to be able to investigate the specificity of multi-hop reasoning skills to particular contexts, we select the following recent synthetic datasets.

GSM-∞ (Zhou et al., 2025) generalizes the GSM8K benchmark of grade school math word problems (Cobbe et al., 2021) to its potentially-infinitely extensible counterpart. GSM-∞ builds a random computation graph to represent the ground-truth solution trace, optionally augmented with distractor facts. It then converts the graph to a word problem via natural language templates in one of the pre-defined themes. This dataset is representative of a common class of math-based synthetic reasoning benchmarks, and due to confounding arithmetic skills it allows us to investigate generalizability of math-based reasoning to more knowledge-intensive real-world tasks.

We generate arithmetic word problems from GSM-∞ using its "medium" difficulty level, set the number of arithmetic operations to range between 2 and 20, and fix the context length to zero: this configuration ensures that each problem only contains information necessary and sufficient for the solution and therefore simplifies the identification of the "hops" in the solution trace. With this synthetic data curation strategy, we collect $\approx$ 600 questions for each of 19 difficulty levels. Half of each set of questions are from *zoo* theme, one quarter from *teacher–school* theme, and the remaining from *movie* theme – all equally split between forward and reverse modes (which correspond to addition/multiplication and subtraction/division questions, respectively). This yields a total of $\approx$ 12.5K samples, with 10K samples used for training and the remainder for validation.

PhantomWiki (Gong et al., 2025) generates on-demand synthetic datasets of natural language document corpora and question answer-pairs, designed to evaluate LLMs on multi-step and multi-branch reasoning and retrieval capabilities. Each PhantomWiki dataset represents a randomly universe of fictional individuals, whose personal attributes and inter-personal relations are described in a set of Wikipedia-like documents; then a context-free grammar and logic programming-based algorithm generates multi-hop reasoning questions like "*Who is the nephew of the friend of the person who likes birdwatching?*". Unlike in GSM-∞, questions in PhantomWiki may have multiple possible answers (*e.g.*, there could be many people who like birdwatching, and each of their friends could have several nephews). It also requires a greater extent of retrieval and knowledge composition skills—finding and processing the relevant information scattered across multiple documents.

In our experiments, we configure our PhantomWiki datasets to only contain "easy" relations like immediate family and friends, so that the "hops" are conceptually simple. We further filter out aggregation questions of the form "*How many ...*", to constrain the dataset to purely multi-hop questions like "*Who is the <relation> of ...?*" and "*What is the <attribute> of ...?*. This setup ensures that answering a question of difficulty d requires hopping through documents of exactly d individuals in the universe and eliminates the confounding counting skill. We generate 34 universes of 25 individuals using 100 random seeds, and set the context-free grammar recursion depth to 20 to generate questions with varying difficulties; we also obtain the ground-truth list of answers for each question using PhantomWiki's logic program. This results in 330 questions per universe with

question difficulties (hops) ranging from 1 to 9. We select 31 universes for 10K training question-answer samples, and reserve 3 universes ($\approx$ 1K samples) for validation.

3.2 RL Fine-tuning for Reasoning

In this work, we use group relative policy optimization (GRPO; Shao et al., 2024) as the primary RL fine-tuning algorithm for understanding reasoning transfer. GRPO has been introduced as a variant of proximal policy optimization (PPO; Schulman et al., 2017); by replacing the value model-based advantage estimation of the original PPO algorithm with one the one based on a *group* of online completions for each prompt, it effectively reduces memory and compute requirements.

For clarity, we restate the objective from Shao et al. (2024): given a question q sampled from a distribution over question set $P(Q)$, GRPO samples a group of G output completions $\{o_1, \dots o_G\}$ from the old LLM π with parameters θ_{old}, and assigns each output completion a scalar reward value $\{R_1, \dots, R_G\}$. The algorithm estimates the advantage $\widehat{A}_i$ of each completion by normalizing with respect to the average reward as a baseline. The final objective is as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\left[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)\right]} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[r_{i,t} \hat{A}_{i,t}, \text{clip}\left(r_{i,t}, 1-\varepsilon, 1+\varepsilon\right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \right\} \right]$$

$$\text{where } \widehat{A}_{i,t} = \frac{R_i - \text{mean}(R_1, \dots, R_G)}{\text{stdev}(R_1, \dots, R_G)}.$$

Here π_{ref} is a reference policy (usually model initialization) used in the KL divergence penalty $\mathbb{D}_{\text{KL}}$, ϵ, β are hyperparameters, and the relative weight $r_{i,t}$ for output completion o_i is calculated on a per-token basis $r_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$.

In our experiments, we use the `GRPOTrainer` implementation from the open-source Hugging Face `TRL` library[1] (von Werra et al., 2020), which implements a per-GPU-device training batch version of GRPO and sets KL-divergence penalty hyperparameter β to 0.

3.3 Prompt and Reward Design

We train LLMs to perform in-context reasoning and retrieval, *i.e.*, to answer questions given *all* the relevant context in the prompt. Our prompt first includes the *evidence*: for a GSM-∞ question, this is the problem statement; for a PhantomWiki question, we provide documents for all 25 individuals of the randomly generated universe. After the evidence, our prompt includes an instruction for the LLM to output the final answer within `<answer>...</answer>` tags, which are the standard output format for DeepSeek-R1 for reasoning questions (Guo et al., 2025a) and subsequent LLMs like the Qwen3 family (Qwen Team, 2025). To further ground the answer output format, we append CoT examples: for GSM-∞, we use 3 of the the automatically-generated CoT ground-truth solutions from the training set; for PhantomWiki we use 11 CoT examples that were originally curated by Gong et al. (2025). Finally, we pose the question to the LLM (our full prompts are included in Appendix C).

We parse model generations using a regular expression to pick the last `<answer>...</answer>` in the generated text, and compare with the ground-truth answer. For GSM-∞ questions, we reward the model with binary reward on the correct numeric value. As PhantomWiki questions can have multiple answers, we reward the model-generated answers with an F1 score between 0 and 1.

3.4 Evaluation Datasets

We evaluate all models on the following 3 in-context question answering datasets that measure multi-hop reasoning capabilities. We randomly subsample 500 from the respective test sets for evaluation. **HotpotQA** (Yang et al., 2018) is a multi-hop question answering dataset containing over 100,000

[1] https://huggingface.co/docs/trl/v0.21.0/grpo_trainer

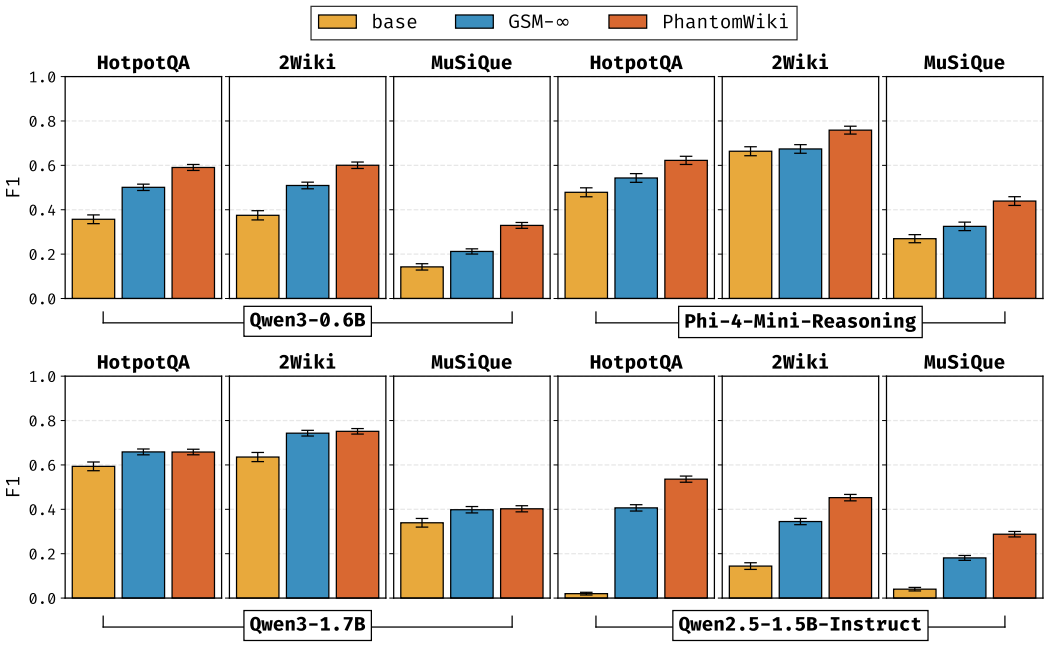

Figure 2: **F1 score on real-world multi-hop reasoning datasets of LLMs finetuned with GRPO on synthetic datasets PhantomWiki and GSM-∞.** We observe that fine-tuning on synthetic reasoning data *consistently* transfers to HotpotQA, 2WikiMultihopQA, and MuSiQue. Concretely, training Qwen3-0.6B model on PhantomWiki improves F1 scores relative to the base model by 62% on HotpotQA, 63% on 2WikiMultihopQA, and 132% on MuSiQue. The performance transfer trends are consistent across model families and sizes (Qwen and Phi LLMs in 1-4B parameter range). We fine-tune each base model with 2 random training seeds, and evaluate final checkpoints of both experiment runs. With this we calculate the standard error, shown as error bars.

questions that require information typically from two Wikipedia paragraphs. Each question follows a consistent two-hop reasoning structure, making it a **2-hop** QA dataset. **2WikiMultihopQA** (Ho et al., 2020) is a more recent **2-hop** dataset, containing over 190,000 two-hop questions organized into four categories: compositional, inference, comparison, and bridge-comparison. The questions are grounded in Wikidata's knowledge graph, with each following a specific two-hop path between related entities. **MuSiQue** (Trivedi et al., 2022) evaluates compositional reasoning with **2-4 hop** questions created by bridging single-hop questions. Questions require joining information from multiple separate paragraphs. We use the MuSiQue-Answerable split of the dataset to ensure that all questions can be answered using a subset of the given context.

4 Results

Performance Transfer from Synthetic to Real-world Datasets. We fine-tune LLMs with GRPO on PhantomWiki and GSM-∞ training datasets, and evaluate their performance on HotpotQA, 2WikiMultihopQA, and MuSiQue. We show in Figure 2 that training on synthetic datasets improves performance across all real-world evaluation datasets. Moreover, this performance transfer is consistent across LLM model families and sizes, with the PhantomWiki dataset transferring better to multi-hop reasoning than GSM-∞. Multi-hop reasoning performance of even Microsoft's Phi-4-mini-reasoning LLM improves with RL fine-tuning, even though the LLM was already trained on synthetic data generated in-house (Abdin et al., 2025).

We separate the model's ability to answer correctly from its ability to answer in the right format with an ablation study. We fine-tune all LLMs for 3K training steps on a binary reward signal for the correct output format in `<answer>...</answer>`: 1 if correct format and 0 if not. In Table 1, we find that Qwen3 family of models and Phi-4-mini-reasoning performance does not improve with format reward training, but Qwen2.5-1.5B-Instruct improves significantly.

		HotpotQA	2WikiMultihopQA	MuSiQue
Qwen3-0.6B	base	0.36 ± 0.02	0.37 ± 0.02	0.14 ± 0.01
	format	0.38 ± 0.02	0.34 ± 0.02	0.13 ± 0.01
Qwen3-1.7B	base	0.59 ± 0.02	0.64 ± 0.02	0.34 ± 0.02
	format	0.64 ± 0.02	0.67 ± 0.02	0.35 ± 0.02
Phi-4-mini-reasoning	base	0.48 ± 0.02	0.66 ± 0.02	0.27 ± 0.02
	format	0.47 ± 0.02	0.48 ± 0.02	0.26 ± 0.02
Qwen2.5-1.5B-Instruct	base	0.02 ± 0.01	0.14 ± 0.02	0.04 ± 0.01
	format	0.43 ± 0.02	0.30 ± 0.02	0.20 ± 0.02

Table 1: **Ablation study on training with binary format reward.** F1 scores of Qwen3 LLMs and Phi-4-mini-reasoning do not improve when trained with binary reward for correct output format within `<answer>...</answer>` (it even hurts in some instances). Qwen2.5-1.5B-Instruct improves remarkably with format reward training. We report standard error on the evaluation datasets.

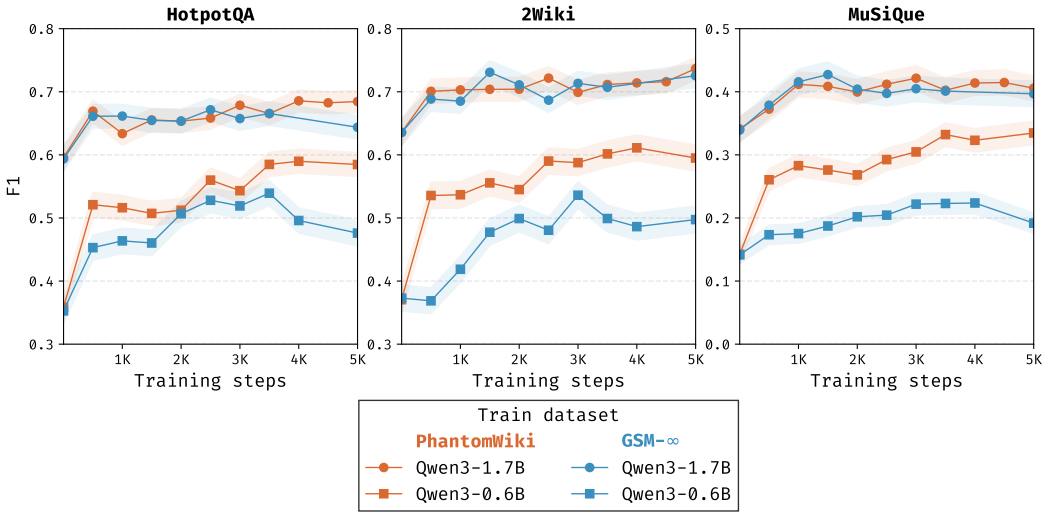

Figure 3: **F1 scores on real-world multi-hop reasoning datasets of intermediate training checkpoints, when LLMs are finetuned with GRPO on synthetic datasets.** We evaluate intermediate checkpoints from every 10% of the full training steps on all evaluation datasets, and show mean $\pm$ standard error with the solid line and shaded region. Performance on all evaluation datasets continues to improve with training steps, especially with PhantomWiki training.

There are two takeaways from this ablation study. First, RL fine-tuning teaches answer formatting, in our case Qwen2.5-1.5B-Instruct LLM. Perhaps this is expected, as the model learns "reward hacking" to elicit the highest reward of 1. Hence, Qwen2.5-1.5B-Instruct model's synthetic-to-real performance transfer in Figure 2 constitutes improvements in both output formatting and correctness. Second and more importantly, RL fine-tuning on synthetic datasets can boost multi-hop reasoning by teaching knowledge composition. This is especially true for the Qwen3 family and Phi-4-mini-reasoning LLMs, which can correctly format outputs at initialization. Since PhantomWiki and GSM-∞ datasets are purely fictional and contain questions that require chaining logical inferences, we attribute *all* improvement in real-world benchmark performance to *knowledge composition*. This demonstrates that LLMs can develop knowledge composition from synthetic data alone, and apply the transferrable skill in real-world settings.

Reasoning Evolves during Training. Knowledge composition requires integrating facts in a chain of logical inferences. So far, we have investigated *what* models learn from synthetic data in RL fine-tuning. This poses a natural question about *how* they learn. To answer this, we evaluate intermediate training checkpoints that are saved at every 10% of the total training steps. Since we train

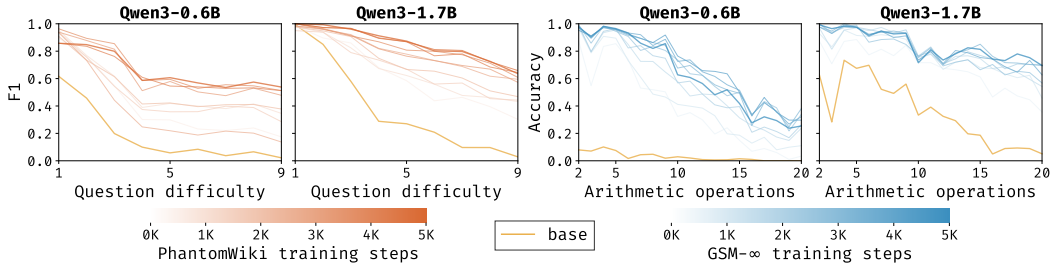

Figure 5: **Reasoning evolution plots of F1 vs question difficulty of intermediate training checkpoints.** We evaluate intermediate training checkpoints of Qwen3-0.6B and Qwen3-1.7B trained on PhantomWiki and GSM-∞ on corresponding validation datasets, and plot the performance as a function of ground-truth question difficulty. On the left are models trained and evaluated on PhantomWiki, stratified by question difficulty that corresponds to necessary number of hops in the PhantomWiki-generated universe. On the right are those on GSM-∞, where the reasoning complexity corresponds to number of arithmetic operations required to answer the math word problem. With continued training and fresh synthetic training samples (lines becoming darker), performance improves on validation questions across all difficulty levels.

for only 1 epoch on the dataset, models see each training sample exactly once. Furthermore, the 10K training samples in PhantomWiki and GSM-∞ are generated from a set of randomly generated universes with no overlap in factual knowledge. This means that evaluating intermediate checkpoints is equivalent to studying the effect of ***synthetic data scaling*** in RL fine-tuning for multi-hop reasoning.

In Figure 3 we find that Qwen3 LLMs continue to improve on real-world multi-hop reasoning benchmarks with more training steps, or equivalently more training samples. This also shows that models do not overfit to the synthetic training dataset. In fact, learning to compose knowledge in fictional worlds of PhantomWiki and GSM-∞ continues to deliver real-world gains. We observe a similar trend for Phi-4-mini-reasoning in Figure 6, although Qwen2.5-1.5B-Instruct does not show such improvement. We note that different LLMs exhibit varying levels of *malleability* for RL fine-tuning: Qwen3-0.6B start off worse but show a steep upward trend while Qwen3-1.7B improve slowly. We leave to future work to analyze how LLM initialization and its "quality" affects RL fine-tuning.

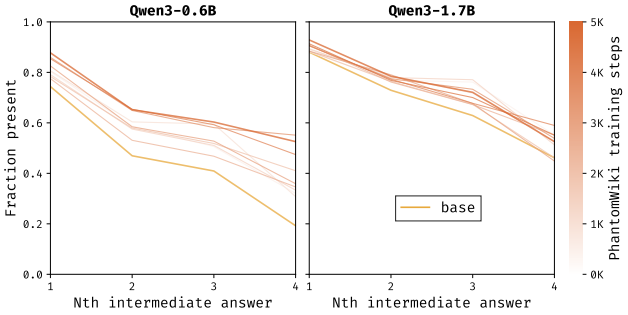

Figure 4: **Reasoning evolution plots on MuSiQue of PhantomWiki training checkpoints.** We evaluate training checkpoints on MuSiQue questions, and plot the fraction of model's generated text that contain the ground-truth n^{th} intermediate answer. With continued training on synthetic data, the LLM reasoning traces include a higher proportion of correct intermediate answers.

Synthetic datasets PhantomWiki and GSM-∞ contain questions of varying difficulties, allowing us to examine model performance as a function of reasoning complexity (Gong et al., 2025; Zhou et al., 2025). Figure 5 decomposes performance of intermediate training checkpoints as a function of question difficulty and arithmetic operations for PhantomWiki and GSM-∞ respectively. The trends are striking: Qwen3 LLMs learn to correctly answer questions across all difficulty levels as training proceeds. Again we see a similar trend for Phi-4-mini-reasoning in Figure 7, and observe Qwen2.5-1.5B-Instruct saturating quickly. Note that universes in validation sets of PhantomWiki and GSM-∞ are completely disjoint from their training sets. Thus, improving on validation questions of all difficulties translates to improving knowledge composition at all levels simultaneously.

Finally, we illustrate that LLMs learn to compose knowledge in real-world MuSiQue dataset in Figure 4. Each question in MuSiQue dataset is supported with a list of ground-truth intermediate answers. On evaluating generations from PhantomWiki training checkpoints, we find that LLMs learn to generate reasoning traces with increasingly higher proportions of intermediate answers. This observation coalesces our findings from performance transfer in Figure 2 and synthetic reasoning evolution in Figure 5 in a key insight: ***the ability to compose knowledge is a fundamental and generalizable skill in multi-hop reasoning tasks, transferring across synthetic and real-world datasets***.

5 Discussion and Future Works

Transferability of Reasoning. Our findings demonstrate that performance on real-world reasoning tasks improves after fine-tuning with synthetic datasets. This cross-domain transfer from fictional to real-world contexts rules out memorization and supports knowledge composition as a transferable meta-skill. This supports that reasoning—specifically chaining logical inferences across multiple steps—constitutes a transferable competency independent of domain-specific factual knowledge (Toplak & Stanovich, 2002). However, the extent of this transferability remains an open question. Real-world reasoning tasks contain both factual knowledge and knowledge composition, and while memorization can degrade performance in counterfactual contexts (Wu et al., 2025a), models can learn memorization and generalizability simultaneously (Xie et al., 2024). Since synthetic datasets are knowledge-free, further investigation of their interplay with knowledge-intensive real-world datasets remains future work.

Synthetic Datasets as Meaningful Training Signals. Beyond simple evaluation, our analysis demonstrates that models learn transferable reasoning capabilities from synthetic datasets. This offers practical advantages: synthetic datasets provide a scalable alternative to human-annotated reasoning data (Villalobos et al., 2024; Muennighoff et al., 2023), positioning domain experts as curators of verifiable curricula. While our work focuses on knowledge composition through multi-hop reasoning, other reasoning capabilities may also transfer via synthetic datasets (Stojanovski et al., 2025). Future work should explore whether causal reasoning, counterfactual inference, or analogical thinking exhibit similar transferability patterns. Furthermore, understanding boundary conditions for synthetic-to-real transfer and extending beyond multi-hop reasoning (Zhao et al., 2023; Wu et al., 2025b; Wang et al., 2024) remain important open questions.

6 Conclusion

In this work, we evaluate the potential of synthetic multi-hop reasoning datasets as a scalable alternative to real-world training data for LLM reasoning. Our results demonstrate that synthetic reasoning training develops transferable compositional inference abilities that achieve significant performance gains on diverse real-world benchmarks, despite zero factual overlap with evaluation domains. This suggests that reasoning transfers across domains, and improvements on synthetic tasks with varying reasoning complexity translates to enhanced performance on real-world reasoning. Our findings demonstrate promising trends towards cost-effective scaling of reasoning capabilities, opening new avenues for developing reasoning-capable language models without traditional data availability constraints.

Ethics Statement

Our work adheres to the ICLR Code of Ethics, and does not pose any societal, personal, or organizational risks.

Reproducibility Statement

To ensure reproducibility, we use free and open-source and software and LLMs. We also include our full dataset preparation, model training, and evaluation configuration in Methodology section and Appendix A. We further report standard errors of all measurements in our results, generate data with fixed random seeds, and set fixed training random seeds where possible.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the Unseen, Logic Reasoning and Degree Curriculum. *Journal of Machine Learning Research*, 25(331):1–58, 2024a. URL http://jmlr.org/papers/v25/24-0220.html.

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How Far Can Transformers Reason? The Globality Barrier and Inductive Scratchpad. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27850–27895. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3107e4bdb658c79053d7ef59cbc804dd-Paper-Conference.pdf.

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, April 2025. URL http://arxiv.org/abs/2504.21318.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB.

Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D. Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating RL for LLM Reasoning with Optimal Advantage Regression, May 2025. URL http://arxiv.org/abs/2505.20686.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward Language Agent Fine-tuning, October 2023. URL http://arxiv.org/abs/2310.05915.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems, May 2025. URL http://arxiv.org/abs/2505.11831. arXiv:2505.11831 [cs].

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL http://jmlr.org/papers/v25/23-0870.html.

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, November 2021. URL http://arxiv.org/abs/2110.14168.

Tri Dao. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

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, August 2025. URL http://arxiv.org/abs/2503.01307.

Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D. Manning. Synthetic Data Generation & Multi-Step RL for Reasoning & Tool Use, April 2025. URL http://arxiv.org/abs/2504.04736.

Albert Gong, Kamilė Stankevičiūtė, Chao Wan, Anmol Kabra, Raphael Thesmar, Johann Lee, Julius Klenke, Carla P. Gomes, and Kilian Q. Weinberger. PhantomWiki: On-Demand Datasets for Reasoning and Retrieval Evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada, June 2025. PMLR. doi: 10.48550/arXiv.2502.20377. URL http://arxiv.org/abs/2502.20377.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. In *Proceedings of the 42nd International Conference on Machine Learning*, Vancouver, Canada, January 2025. URL https://openreview.net/forum?id=5zwF1GizFa.

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, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, 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, Mingxu Zhou, 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, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, 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, 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 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025a. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I. Jordan, and Stuart Russell. How Do LLMs Perform Two-Hop Reasoning in Context?, May 2025b. URL http://arxiv.org/abs/2502.13913.

Seungwook Han, Jyothish Pari, Samuel J. Gershman, and Pulkit Agrawal. Position: General Intelligence Requires Reward-based Pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, June 2025. doi: 10.48550/arXiv.2502.19402. URL http://arxiv.org/abs/2502.19402.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E. Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space. In *Workshop on Reasoning and Planning for Large Language Models at ICLR 2025*, 2025. URL https://openreview.net/forum?id=KrWSrrYGpT.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, December 2021. doi: 10.48550/arXiv.2103.03874. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://www.aclweb.org/anthology/2020.coling-main.580.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: An Efficient RLHF Algorithm with Robustness to Both Prompt and Reward Models, August 2025. URL http://arxiv.org/abs/2501.03262.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, and Lichao Sun. DataGen: Unified Synthetic Dataset Generation via Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=F5R0lG74Tu.

Alihan Hüyük, Xinnuo Xu, Jacqueline R. M. A. Maasch, Aditya V. Nori, and Javier Gonzalez. Reasoning Elicitation in Language Models via Counterfactual Feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VVixJ9QavY.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 979-8-4007-0229-7. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165. event-place: Koblenz, Germany.

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. Tülu 3: Pushing Frontiers in Open Language Model Post-Training, April 2025. URL http://arxiv.org/abs/2411.15124.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and Xian Li. NaturalThoughts: Selecting and Distilling Reasoning Traces for General Reasoning Tasks, July 2025. URL http://arxiv.org/abs/2507.01921.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. March 2025. doi: 10.48550/arXiv.2503.20783. URL http://arxiv.org/abs/2503.20783. arXiv:2503.20783 [cs].

MAA. MAA invitational competitions. american invitational mathematics examination (aime). URL https://maa.org/maa-invitational-competitions/.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, Singapore, March 2025. doi: 10.48550/arXiv.2410.05229. URL https://openreview.net/forum?id=AjXkRZIvjB.

David Mizrahi, Anders Boesen Lindbo Larsen, Jesse Allardice, Suzie Petryk, Yuri Gorokhov, Jeffrey Li, Alex Fang, Josh Gardner, Tom Gunter, and Afshin Dehghan. Language Models Improve When Pretraining Data Matches Target Tasks, July 2025. URL http://arxiv.org/abs/2507.12466.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling Data-Constrained Language Models. 36:50358–50376, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative Reasoning Preference Optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 116617–116637. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d37c9ad425fe5b65304d500c6edcba00-Paper-Conference.pdf.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large Language Model Connected with Massive APIs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=tBRNC6YemY.

Peng Qi, Haejun Lee, Tg Sido, and Christopher Manning. Answering Open-Domain Questions of Varying Reasoning Steps from Text. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3599–3614, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.292. URL https://aclanthology.org/2021.emnlp-main.292/.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*, 2025.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. ARB: Advanced Reasoning Benchmark for Large Language Models. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS 2023*. arXiv, July 2023. doi: 10.48550/arXiv.2307.13692. URL http://arxiv.org/abs/2307.13692.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017. URL http://arxiv.org/abs/1707.06347.

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. June 2025. doi: 10.48550/arXiv.2506.10947. URL http://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, April 2024. URL http://arxiv.org/abs/2402.03300.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity, July 2025. URL http://arxiv.org/abs/2506.06941.

Vaishnavi Shrivastava, Ahmed Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. *arXiv preprint arXiv:2508.09726*, 2025.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458/.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. REASONING GYM: Reasoning Environments for Reinforcement Learning with Verifiable Rewards, May 2025. URL http://arxiv.org/abs/2505.24760. arXiv:2505.24760 [cs].

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*, pp. 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL http://aclweb.org/anthology/N19-1421.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=t4eB3zYWBK.

Maggie E Toplak and Keith E Stanovich. The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of educational psychology*, 94(1):197, 2002.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, January 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06747-5. URL https://www.nature.com/articles/s41586-023-06747-5.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, May 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00475. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00475/110996/MuSiQue-Multihop-Questions-via-Single-hop-Question.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. 2024. URL https://openreview.net/forum?id=ViZcgDQjyG.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Chao Wan, Albert Gong, Mihir Mishra, Carl-Leander Henneking, Claas Beger, and Kilian Q Weinberger. Memento: Note-Taking for Your Future Self. June 2025. URL https://arxiv.org/abs/2506.20642.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs, June 2025. URL http://arxiv.org/abs/2506.14245.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. CofCA: A Step-Wise Counterfactual Multi-hop QA benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=q2DmkZ1wVe.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or Memorization? Unreliable Results of Reinforcement Learning Due to Data Contamination, July 2025b. URL http://arxiv.org/abs/2507.10532.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. Towards System 2 Reasoning in LLMs: Learning How to Think With Meta Chain-of-Thought, January 2025. URL http://arxiv.org/abs/2501.04682. arXiv:2501.04682 [cs].

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On Memorization of Large Language Models in Logical Reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*. arXiv, 2024. doi: 10.48550/arXiv.2410.23123. URL https://openreview.net/forum?id=mxX8WdPCx9.

Roy Xie, David Qiu, Deepak Gopinath, Dong Lin, Yanchao Sun, Chong Wang, Saloni Potdar, and Bhuwan Dhingra. Interleaved Reasoning for Large Language Models via Reinforcement Learning, May 2025. URL http://arxiv.org/abs/2505.19640.

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

Processing, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL http://aclweb.org/anthology/D18-1259.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=07yvxWDSla.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. $\backslashtau $-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045*, 2024.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. SKILL-MIX: a Flexible and Expandable Family of Evaluations for AI Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Jf5gplvglq.

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. May 2025. doi: 10.48550/arXiv.2503.14476. URL http://arxiv.org/abs/2503.14476. arXiv:2503.14476 [cs].

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?, May 2025. URL http://arxiv.org/abs/2504.13837.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping Reasoning With Reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander Rush. Abductive Commonsense Reasoning Exploiting Mutually Exclusive Explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14883–14896, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.831. URL https://aclanthology.org/2023.acl-long.831.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to Reason without External Rewards, August 2025. URL http://arxiv.org/abs/2505.19590. arXiv:2505.19590 [cs].

Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. GSM-Infinite: How Do Your LLMs Behave over Infinitely Increasing Context Length and Reasoning Complexity? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, Vancouver, Canada, February 2025. PMLR. URL https://openreview.net/forum?id=n52yyvEwPa.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. ToolQA: A Dataset for LLM Question Answering with External Tools. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50117–50143. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/

9cb2a7495900f8b602cb10159246a016-Paper-Datasets_and_Benchmarks.pdf.

A Experiment Configuration

We use the 0.21.0 version of Hugging Face TRL library's `GRPOTrainer` using vLLM colocate mode (Kwon et al., 2023) and FlashAttention-2 (Dao, 2024) on 4 NVIDIA H100 GPUs each with 80GB VRAM. With the configuration in Listing 1, RL finetuning a 1 to 4B parameter LLM on 10K training samples takes ≈1 day on our Linux cluster. [2] Since Phi-4-mini-reasoning is a 4B parameter LLM, we adjust the `vllm_gpu_memory_utilization: 0.25`, `per_device_train_batch_size: 4`, and `num_generations: 8` to train on 4 H100 GPUs each with 80GB VRAM. The prompt length varies for each training dataset, and we adjust the `max_prompt_length` to prevent prompt truncation:

1. PhantomWiki: 6000
2. GSM-∞: 2048
3. HotpotQA: 6000
4. 2WikiMultihopQA: 6000
5. MuSiQue: 8000

```yaml
# Training parameters
per_device_train_batch_size: 8
gradient_accumulation_steps: 1
num_generations: 16

# vLLM settings
use_vllm: true
vllm_mode: "colocate"
vllm_gpu_memory_utilization: 0.20

# Generation parameters
max_completion_length: 4096
temperature: 1.0
top_p: 1.0
top_k: null
min_p: null
repetition_penalty: 1.0

# GRPO algorithm parameters
beta: 0.0
epsilon: 0.2
importance_sampling_level: "token"
scale_rewards: true
loss_type: bnpo
mask_truncated_completions: false
```

Listing 1: GRPOTrainer hyperparameter values in our YAML configuration file

B Additional Results

[2] The larger models Qwen3-1.7B and Phi-4-mini-reasoning take the full 1 day, i.e. using ≈100 H100 hours per training experiment as they generate long CoT. The Qwen2.5-1.5B-Instruct model does not generate long CoT, and thus trains the fastest in ≈20 H100 hours.

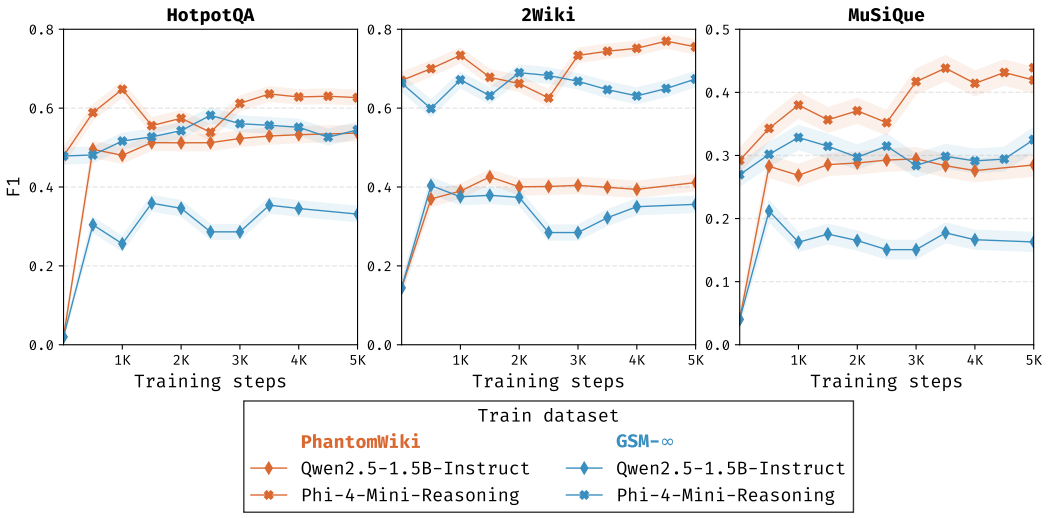

Figure 6: **F1 scores on real-world multi-hop reasoning datasets of intermediate training checkpoints, when LLMs are finetuned with GRPO on synthetic datasets.** We evaluate intermediate checkpoints from every 10% of the full training steps on all evaluation datasets, and show mean ± standard error with the solid line and shaded region. Performance on all evaluation datasets generally improves with training steps for Phi-4-mini-reasoning, but saturates for Qwen2.5-1.5B-Instruct.

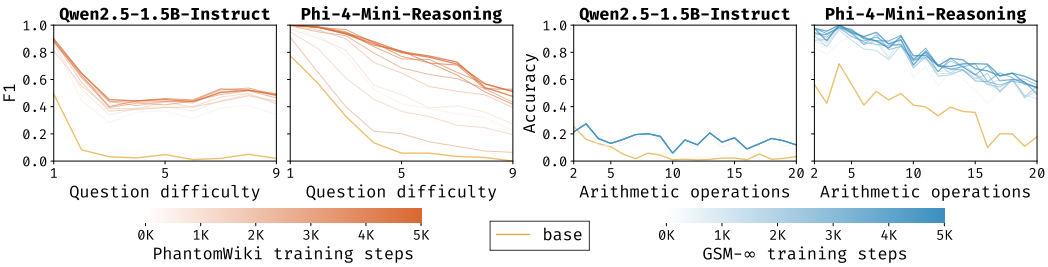

Figure 7: **Reasoning evolution plots of F1 vs question difficulty of intermediate training checkpoints.** We evaluate intermediate training checkpoints of Qwen2.5-1.5B-Instruct and Phi-4-mini-reasoning trained on PhantomWiki and GSM-∞ on corresponding validation datasets, and plot the performance as a function of ground-truth question difficulty. On the left are models trained and evaluated on PhantomWiki, decomposed on question difficulty that corresponds to necessary number of hops in the PhantomWiki-generated universe. On the right are those on GSM-∞, where the reasoning complexity corresponds to number of arithmetic operations required to answer the math word problem. With continued training and fresh synthetic training samples (lines becoming darker), performance improves on validation questions across all difficulty levels. Qwen2.5-1.5B-Instruct saturates quickly on GSM-∞—the reason why the final checkpoint in dark blue hides the intermediate checkpoint lines in the right plot.

C Prompts

C.1 PhantomWiki prompt

We use CoT prompt template and examples from (Gong et al., 2025), with a custom instruction asking the LLM to output the final answer within `<answer>...</answer>`.

```
You are given the following evidence:
(BEGIN EVIDENCE)
{{evidence}}
(END EVIDENCE)
```

```
You will be provided a question. Your response must end with the
    final answer enclosed in tags: <answer>FINAL_ANSWER</answer>

Here, FINAL_ANSWER must be one of the following:
- a name (if there is only one correct answer);
- a list of names separated by ',' (if there are multiple correct
    answers); or
- numbers separated by ',' (if the answer is numerical); or
- empty string (if there is no answer).

Here are some examples:
(START OF EXAMPLES)
Example 1:
Question: Who is the sister of Aida Wang?
Answer: Based on the evidence, the sisters of Aida Wang are
    Barabara Beltran, Vicki Hackworth. <answer>Barabara Beltran,
    Vicki Hackworth</answer>.

Example 2:
Question: Who is the child of Alvaro Smock?
Answer: Based on the evidence, the children of Alvaro Smock are
    Eli Smock, Gene Smock. <answer>Eli Smock,Gene Smock</answer>.

Example 3:
Question: Who is the friend of the child of Alvaro Smock?
Answer: First I need to find the children of Alvaro Smock. Based
    on the evidence, the children of Alvaro Smock are Eli Smock,
    Gene Smock. Now I need to find the friends of Eli Smock and
    Gene Smock. Based on the evidence, the friends of Eli Smock
    are Leisa Lutz, Shelli Beltran, Vicki Hackworth, Virgil
    Hackworth, Alison Smock, Brian Beltran. The friends of Gene
    Smock are Leeann Hackworth, Leisa Lutz, Ricardo Hackworth,
    Alvaro Smock, Dominique Smock. <answer>Leisa Lutz,Shelli
    Beltran,Vicki Hackworth,Virgil Hackworth,Alison Smock,Brian
    Beltran,Leeann Hackworth,Ricardo Hackworth,Dominique Smock</
    answer>.

Example 4:
Question: Who is the aunt of Vicki Hackworth?
Answer: An aunt is the sister of a parent. Based on the evidence,
    the parents of Vicki Hackworth are Shelli Beltran, Dino
    Beltran. To find the aunt of Vicki Hackworth, I need to find
    the sister of Shelli Beltran and Dino Beltran. Based on the
    evidence, Shelli Beltran has no sister, and the sister of Dino
     Beltran is Stacia Toombs. <answer>Stacia Toombs</answer>.

Example 5:
Question: What is the occupation of the husband of Stacia Toombs?
Answer: Based on the evidence, the husband of Stacia Toombs is
    Wilbert Toombs. The occupation of Wilbert Toombs is theatre
    manager. <answer>theatre manager</answer>.

Example 6:
Question: What is the hobby of the daughter-in-law of Lannie Smock
    ?
Answer: A daughter-in-law is the wife of a child. Based on the
    evidence, the children of Lannie Smock are Eli Smock, Gene
    Smock. Eli Smock has no wife, and the wife of Gene Smock is
```

Dominique Smock. The hobby of Dominique Smock is dominoes. <answer>dominoes</answer>.

Example 7:
Question: What is the date of birth of the person whose hobby is finance?
Answer: I need to search for people whose hobby is finance. Based on the evidence, the person whose hobby is finance is Stacia Toombs. The date of birth of Stacia Toombs is 0959-03-22. <answer>0959-03-22</answer>.

Example 8:
Question: Who is the great-granddaughter of the person whose occupation is biomedical scientist?
Answer: I need to search for people whose occupation is biomedical scientist. Based on the evidence, the person whose occupation is biomedical scientist is Lannie Smock. To find the great-granddaughter of Lannie Smock, I need to find the daughter of the child of the child of Lannie Smock. Based on the evidence, the children of Lannie Smock are Eli Smock, Gene Smock. Eli Smock has no child, and the child of Gene Smock is Williams Smock. The daughters of Williams Smock are Shelli Beltran, Stacia Toombs. <answer>Shelli Beltran,Stacia Toombs</answer>.

Example 9:
Question: How many friends does Ryan Wang have?
Answer: Based on the evidence, the friends of Ryan Wang are Shelli Beltran, Stacia Toombs, Virgil Hackworth, Aida Wang. <answer>4</answer>.

Example 10:
Question: How many friends does the child of Alvaro Smock have?
Answer: First, I need to find the children of Alvaro Smock. Based on the evidence, the children of Alvaro Smock are Eli Smock, Gene Smock. Now I need to find how many friends they have. Based on the evidence, the friends of Eli Smock are Leisa Lutz, Shelli Beltran, Vicki Hackworth, Virgil Hackworth, Alison Smock, Brian Beltran. The friends of Gene Smock are Leeann Hackworth, Leisa Lutz, Ricardo Hackworth, Alvaro Smock, Dominique Smock. <answer>6,5</answer>.

Example 11:
Question: How many uncles does the friend of Stacia Toombs have?
Answer: First, I need to find the friends of Stacia Toombs. Based on the evidence, the friends of Stacia Toombs are Brian Beltran, Isiah Lutz, Leeann Hackworth, Lesley Lutz, Ryan Wang. Now I need to find how many uncles they have. An uncle is the brother of a parent. Based on the evidence, Brian Beltran has no parents, Isiah Lutz has no parents, Leeann Hackworth has 2 parents, Lesley Lutz has 2 parents, and Ryan Wang has no parents. Based on the evidence, the parents of Leeann Hackworth are Vicki Hackworth, Ricardo Hackworth. But both parents do not have brothers. Based on the evidence, the parents of Lesley Lutz are Leisa Lutz, Isiah Lutz. The brother of Leisa Lutz is Virgil Hackworth, so he is an uncle of Lesley Lutz. Isiah Lutz has no brother. So the friends of Stacia Toombs have 0, 0, 0, 1, 0 uncles. Unique is 0, 1. <answer>0,1</answer>.
(END OF EXAMPLES)

```
Question: {{question}}
Answer: """
```

C.2 GSM-∞ PROMPT

We modify the CoT prompt template from PhantomWiki (Gong et al., 2025) by replacing `EVIDENCE` with the problem statement. GSM-∞ also generates a templated solution for each question pairs, which we use as the CoT examples in the prompt.

```
You are given the following problem:
(BEGIN PROBLEM)
{{problem}}
(END PROBLEM)

You will be provided a question on the above problem. Your
    response must end with the final answer enclosed in tags: <
    answer>FINAL_ANSWER</answer>

Here, FINAL_ANSWER must be a number.

Here are some examples:
(START OF EXAMPLES)
Example 1:
Question: What is the total number of adult animals in Maple Creek
    ?
Answer: Define adult wolf in Maple Creek as r; so r = 2. Define
    total number of adult animals in Maple Creek as p; so p = r =
    2. <answer>2</answer>.

Example 2:
Question: What is the total number of schools in Clearwater Bay?
Answer: Define elementary school in Riverton City as b; so b = 3.
    Define private middle school in Clearwater Bay as i; so i = b
    = 3. Define public highschool in Clearwater Bay as M; so M = i
     = 3. Define elementary school in Clearwater Bay as G; so G =
    2. Define total number of schools in Clearwater Bay as W; V =
    G + i = 2 + 3 = 5; so W = V + M = 5 + 3 = 8. <answer>8</answer
    >.

Example 3:
Question: What is the total number of movies in Festival de
    Clairmont?
Answer: Define upbeat metropolis comedy in Festival de Saint-
    Rivage as m; so m = 4. Define total number of movies in
    Festival de Saint-Rivage as k; so k = m = 4. Define intense
    detective thriller in Festival Lumi\u00e8re de Valmont as C; l
     = k - m = 4 - 4 = 0; so C = 3 + l = 3 + 0 = 3. Define total
    number of movies in Festival Lumi\u00e8re de Valmont as Q; so
    Q = C = 3. Define solemn period drama in R\u00eaves de
    Belleville as N; t = Q + C = 3 + 3 = 6; T = t + k = 6 + 4 =
    10; so N = 4 + T = 4 + 10 = 14. Define total number of movies
    in R\u00eaves de Belleville as y; so y = N = 14. Define
    futuristic sci-fi movie in Festival de Clairmont as A; z = y +
     N = 14 + 14 = 28; q = z + C = 28 + 3 = 31; so A = 3 * q = 3 *
     31 = 93. Define total number of movies in Festival de
    Clairmont as p; so p = A = 93. <answer>93</answer>.
(END OF EXAMPLES)
```

```
Question: {{question}}
Answer:
```

LLM USE

LLMs were used to revise and proofread paper content. All claims have been verified and cross-referenced by the authors.