Novelty Assessment Report

Paper: QuestA: Expanding Reasoning Capacity in LLMs via Question Augmentation
PDF URL: https://openreview.net/pdf?id=3MifB0f7qR
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Reinforcement learning (RL) has emerged as a central paradigm for training large language models (LLMs) in reasoning tasks. Yet recent studies question RL's ability to incentivize reasoning capacity beyond the base model. This raises a key challenge: how can RL be adapted to solve harder reasoning problems more effectively? To address this challenge, we propose a simple yet effective strategy via Question Augmentation: introduce partial solutions during training to reduce problem difficulty and provide more informative learning signals. Our method, QuestA, when applied during RL training on math reasoning tasks, not only improves pass@1 but also pass@k—particularly on problems where standard RL struggles to make progress. This enables continual improvement over strong open-source models such as DeepScaleR and OpenMath Nemotron, further enhancing their reasoning capabilities. We achieve new state-of-the-art results on math benchmarks using 1.5B-parameter models: 72.50\% (+10.73\%) on AIME24, 62.29\% (+12.79\%) on AIME25, and 41.67\% (+10.11\%) on HMMT25. Code, data and model are available at https://anonymous.4open.science/r/questa932.

Core Task Landscape

This paper addresses: **Enhancing Reasoning Capacity in Large Language Models through Reinforcement Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **16 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core RL Algorithms and Training Methods**
- **Empirical Analysis of RL for Reasoning**
- **Reasoning Paradigms and Architectures**
- **Domain-Specific Applications**
- **Resource-Constrained and Efficient Methods**
- **Alternative Paradigms and Architectures**
- **External Knowledge and Grounding**
- **Survey and Review Papers**

Complete Taxonomy Tree

- Enhancing Reasoning Capacity in Large Language Models through Reinforcement Learning Survey Taxonomy
- Core RL Algorithms and Training Methods
  - Policy Optimization Algorithms (4 papers)
  - [9] Teaching Large Language Models to Reason with Reinforcement Learning (Havrilla, 2024) View paper
  - [14] Effective Reinforcement Learning for Reasoning in Language Models (Li Shuo, 2025) View paper
  - [19] Dapo: An open-source llm reinforcement learning system at scale (Yu, 2025) View paper
  - [45] Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment (Amirhossein Kazemnejad, 2024) View paper
  - Reward Design and Credit Assignment (3 papers)
  - [12] Rethinking Reasoning Quality in Large Language Models through Enhanced Chain-of-Thought via RL (He, 2025) View paper
  - [24] Offline reinforcement learning for llm multi-step reasoning (Bao Yilin, 2025) View paper
  - [50] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning (Setlur, 2024) View paper
  - Training Strategies and Optimization (4 papers)
  - [7] Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning (Xie Tian, 2025) View paper
  - [21] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't (Quy-Anh Dang, 2025) View paper
  - [33] Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning (Xi, 2024) View paper
  - [42] Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning (Zhou Yang, 2025) View paper
- Empirical Analysis of RL for Reasoning
  - Capability Emergence and Boundaries (4 papers)
  - [1] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? (Yue Yang, 2025) View paper
  - [2] Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models (Liu Mingjie, 2025) View paper
  - [5] Reinforcement learning for reasoning in large language models with one training example (Wang Yi-ping, 2025) View paper
  - [26] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms (Liu Zihan, 2025) View paper
  - Scaling Laws and Training Dynamics (2 papers)
  - [16] Demystifying long chain-of-thought reasoning in llms (Tong, 2025) View paper

    - [44] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning (Tan Ze-lin, 2025) View paper
- Reasoning Paradigms and Architectures
    - Chain-of-Thought and Reasoning Traces (4 papers)
    - [3] Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (DeepSeek-AI, 2025) View paper
    - [29] DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning (Daya Guo, 2025) View paper
    - [32] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning (Ning, 2025) View paper
    - [36] Interleaved Reasoning for Large Language Models via Reinforcement Learning (Qiu, 2025) View paper
    - Search and Exploration Integration (4 papers)
    - [4] Learning to reason with search for llms via reinforcement learning (M Chen, 2025) View paper
    - [25] Advancing language model reasoning through reinforcement learning and inference scaling (Hou Zhenyu, 2025) View paper
    - [28] Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search (Shen, 2025) View paper
    - [39] Reasoning with Exploration: An Entropy Perspective (Cheng, 2025) View paper
    - Multi-Turn and Collaborative Reasoning (2 papers)
    - [47] An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents (Jin Bo-wen, 2025) View paper
    - [49] Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks (Zhou, 2025) View paper
- Domain-Specific Applications
    - Mathematical Reasoning ★ (3 papers)
    - [0] QuestA: Expanding Reasoning Capacity in LLMs via Question Augmentation (Anon et al., 2026) View paper
    - [18] WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct (Luo, 2023) View paper
    - [48] Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning (Zhang Zi-yin, 2025) View paper
    - Software Engineering and Coding (1 papers)
    - [37] Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution (Wei, 2025) View paper
    - Multimodal and Vision-Language Reasoning (3 papers)
    - [11] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models (Guanghao Zhou, 2025) View paper
    - [13] Fine-tuning large vision-language models as decision-making agents via reinforcement learning (Hao Bai, 2024) View paper
    - [22] Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning (Wan, 2025) View paper
    - Specialized Applications (4 papers)
    - [10] Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning (Zhuang, 2025) View paper
    - [38] Reinforced Latent Reasoning for LLM-based Recommendation (Zhang Yang, 2025) View paper
    - [43] A Collaborative Reasoning Framework Powered by Reinforcement Learning and Large Language Models for Complex Questions Answering over Knowledge Graph (Z Zhang, 2025) View paper
    - [46] Eliciting Chain-of-Thought Reasoning for Time Series Analysis using Reinforcement Learning (Parker, 2025) View paper
- Resource-Constrained and Efficient Methods (2 papers)
    - [34] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router (Shao, 2025) View paper
    - [40] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models (Le, 2025) View paper
- Alternative Paradigms and Architectures (1 papers)
    - [31] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning (Zhao Siyan, 2025) View paper
- External Knowledge and Grounding (3 papers)
    - [17] Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning (Carta, 2023) View paper
    - [23] Enhance Reasoning for Large Language Models in the Game Werewolf (Wu Shuang, 2024) View paper
    - [41] Guiding Pretraining in Reinforcement Learning with Large Language Models (Du Yuqing, 2023) View paper
- Survey and Review Papers (7 papers)
    - [6] Advancing reasoning in large language models: Promising methods and approaches (Patil Avinash, 2025) View paper
    - [8] Towards large reasoning models: A survey of reinforced reasoning with large language models (Xu Fengli, 2025) View paper
    - [15] A survey of reinforcement learning for large reasoning models (Zhang Kai-Yan, 2025) View paper
    - [20] A survey on large language models for mathematical reasoning (Wang Peng-Yuan, 2025) View paper
    - [27] A technical survey of reinforcement learning techniques for large language models (Aggarwal, 2025) View paper
    - [30] Multi-step reasoning with large language models, a survey (Aske Plaat, 2025) View paper
    - [35] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle (Liu Ke-liang, 2025) View paper

Narrative

Core task: Enhancing reasoning capacity in large language models through reinforcement learning. The field has organized itself into several major branches that reflect both methodological diversity and application focus. At the highest level, researchers distinguish between core RL algorithms and training methods—where foundational techniques such as policy gradient variants and reward modeling are developed—and empirical analyses that systematically evaluate how RL shapes reasoning behavior. Parallel branches address reasoning paradigms and architectures (exploring chain-of-thought, search-based inference, and latent reasoning structures), domain-specific applications (including mathematical reasoning, code generation, and interactive agents), and resource-constrained methods that prioritize efficiency. Additional branches cover alternative paradigms (such as diffusion-based or non-standard architectures), external knowledge integration, and survey papers that synthesize emerging trends. Works like Deepseek-r1[3] and Large Reasoning Models[8] illustrate how core RL techniques scale to complex reasoning tasks, while WizardMath[18] and Deeptheorem[48] exemplify domain-specific mathematical applications.

Within this landscape, mathematical reasoning has emerged as a particularly active testbed, drawing on both supervised fine-tuning pipelines and RL-driven exploration to handle multi-step problem solving. Many studies in this branch investigate trade-offs between sample efficiency, verifiability of intermediate steps, and the balance between exploration and exploitation during training. QuestA[0] situates itself squarely in this mathematical reasoning cluster, emphasizing RL-based enhancement of reasoning chains for question-answering tasks. Its approach aligns closely with neighbors like WizardMath[18], which also targets mathematical problem solving, and Deeptheorem[48], which extends RL methods to formal theorem proving. Compared to these works, QuestA[0] appears to focus on refining the reward signal and training dynamics specific to question-driven reasoning, contributing to ongoing efforts to make RL more effective and interpretable in structured mathematical domains.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct

Authors: Luo, Haipeng, Sun, Qingfeng, Haipeng Luo, et al. (28 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Large language models (LLMs), such as GPT-4, have shown remarkable performance in natural language processing (NLP) tasks, including challenging mathematical reasoning. However, most existing open-source models are only pre-trained on large-scale internet data and without math-related optimization. In this paper, we present WizardMath, which enhances the mathematical CoT reasoning abilities of LLMs without using external python tools, by applying our proposed Reinforcement Learning from Evol-Ins...

Relationship Analysis

Both papers belong to the Mathematical Reasoning category, focusing on enhancing LLMs' mathematical problem-solving capabilities through reinforcement learning. While QuestA addresses the challenge of training on hard problems by augmenting questions with partial solutions to improve RL efficiency, WizardMath takes a different approach by evolving mathematical instructions through upward and downward difficulty adjustments and combining instruction-level and process-level reward models (IRM and PRM) for reinforcement learning. The key distinction is that QuestA focuses on curriculum learning through partial solution hints during training, whereas WizardMath emphasizes data augmentation through instruction evolution and dual reward modeling.

---

2. Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning

Authors: Zhang Zi-yin, Xu Jiahao, Ziyin Zhang, He Zhiwei, Jiahao Xu, et al. (30 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Theorem proving serves as a major testbed for evaluating complex reasoning abilities in large language models (LLMs). However, traditional automated theorem proving (ATP) approaches rely heavily on formal proof systems that poorly align with LLMs' strength derived from informal, natural language knowledge acquired during pre-training. In this work, we propose DeepTheorem, a comprehensive informal theorem-proving framework exploiting natural language to enhance LLM mathematical reasoning. DeepThe...

Relationship Analysis

Both papers belong to the Mathematical Reasoning category, focusing on RL-based approaches to enhance mathematical problem-solving in LLMs. They overlap in applying reinforcement learning to improve reasoning capacity on challenging math tasks, with both using GRPO and addressing the challenge of sparse rewards on hard problems. However, QuestA focuses on question augmentation via partial solutions to scaffold RL training on existing math datasets, while DeepTheorem introduces a large-scale informal theorem-proving dataset (121K theorems) with a novel RL-Zero training paradigm specifically designed for theorem proving rather than general math problem-solving.

Contributions Analysis

Overall novelty summary. The paper proposes QuestA, a question augmentation strategy that introduces partial solutions during RL training to reduce problem difficulty and improve learning signals for mathematical reasoning. It sits within the Mathematical Reasoning leaf of the taxonomy, which contains only three papers total, indicating a relatively focused but not overcrowded research direction. The sibling papers—WizardMath and Deeptheorem—similarly target RL-based mathematical problem solving, suggesting this work occupies a well-defined niche within domain-specific applications of RL for reasoning enhancement.

The taxonomy reveals that Mathematical Reasoning is one subcategory under Domain-Specific Applications, alongside Software Engineering, Multimodal Reasoning, and Specialized Applications. Neighboring branches include Core RL Algorithms (policy optimization, reward design) and Reasoning Paradigms (chain-of-thought, search integration). QuestA's focus on question augmentation connects it to Training Strategies and Optimization under Core RL Algorithms, as partial solutions can be viewed as a curriculum learning or data augmentation technique. The taxonomy's scope notes clarify that domain-specific methods like QuestA are distinguished from general-purpose algorithm design, positioning this work as an application-driven adaptation rather than a foundational RL innovation.

Among 29 candidates examined, the contribution-level analysis reveals mixed novelty signals. The QuestA method itself examined 9 candidates with 1 refutable match, suggesting some prior work on question augmentation or partial-solution strategies exists within the limited search scope. The theoretical analysis examined 10 candidates with 2 refutable matches, indicating that benefits of partial-solution augmentation may have been explored previously. The state-of-the-art results contribution examined 10 candidates with 1 refutable match, implying that performance claims on these benchmarks face some prior competition. These statistics reflect a top-K semantic search, not an exhaustive literature review, so the presence of refutable candidates indicates overlap within the examined subset rather than definitive lack of novelty.

Given the limited search scope of 29 candidates, the analysis suggests QuestA introduces a focused adaptation of RL training for mathematical reasoning, with some overlap in each contribution area among the examined papers. The work appears to refine existing ideas—question augmentation, partial solutions, and benchmark performance—rather than introduce entirely unprecedented concepts. However, the sparse Mathematical Reasoning leaf and the specific combination of techniques may still offer incremental value. A broader literature search would be needed to assess whether the integration of these elements constitutes a meaningful advance over the field's current state.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: QuestA method via question augmentation with partial solutions

Description: The authors introduce QuestA, a data augmentation approach that prepends partial solutions to hard reasoning problems during reinforcement learning training. This method scaffolds difficult problems by revealing intermediate steps, making them more tractable while providing denser reward signals for more efficient RL training.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Generative question refinement with deep reinforcement learning in retrieval-based QA system

URL: View paper

Brief Assessment

Question Refinement[56] focuses on refining ill-formed user questions in retrieval-based QA systems using seq2seq models with RL rewards based on answer correlation. QuestA augments hard reasoning problems with partial solutions during RL training to scaffold

difficulty and provide denser reward signals for mathematical reasoning tasks. These are fundamentally different approaches serving different purposes.

---

2. Mixture of Autoencoder Experts Guidance using Unlabeled and Incomplete Data for Exploration in Reinforcement Learning

URL: View paper

Brief Assessment

Autoencoder Experts[55] focuses on using unlabeled and incomplete expert demonstrations to guide exploration in continuous control tasks (e.g., MuJoCo robotics), not on augmenting reasoning problems with partial solutions for language model training.

---

3. From Data-Centric to Sample-Centric: Enhancing LLM Reasoning via Progressive Optimization

URL: View paper

Prior Art Analysis

Sample-Centric[58] demonstrates prior work on using partial solution prefixes to guide reinforcement learning training. Both papers propose augmenting training prompts with partial expert solutions to improve RL training on difficult problems. Sample-Centric[58] introduces 'prefix-guided sampling' that appends partial solution prefixes from expert models to challenging problems, which is functionally equivalent to QuestA's approach of prepending partial solutions to hard reasoning problems. The candidate paper explicitly describes this as a data augmentation technique that 'incorporates partial solution prefixes from expert demonstrations to guide the policy, particularly for challenging instances,' directly paralleling QuestA's core mechanism.

Evidence

Evidence 1 - **Rationale**: Both papers describe augmenting difficult problems with partial solutions during RL training. Sample-Centric[58]'s 'prefix-guided sampling' directly corresponds to QuestA's method of augmenting questions with partial solutions. - **Original**: we propose quest a: a parsimonious and efficient strategy that dynamically adjusts problem difficulty during rl training. the core contributions of this work are threefold: • we notice that the evolution of model capacity in rlvr critically depends on dataset difficulty, underscoring the importance ... - **Candidate**: we propose prefix-guided sampling (pg-sampling). this data augmentation strategy is inspired by partially observable reasoning trajectories. the technique involves guiding the policy by providing partial solutions as hints for difficult training samples. these prefixes are sampled from successful so...

Evidence 2 - **Rationale**: Both papers describe the technical mechanism of extracting partial solutions and prepending them to prompts. Sample-Centric[58] provides the mathematical formulation for determining prefix length, showing this is a well-developed prior approach. - **Original**: quest a is a modular augmentation framework designed to inject partial solution sketches into prompts during reinforcement learning (rl) training. it adresses scenarios where the base model fails to generate correct completions-conditions that typically result in sparse reward signals. distinct from... - **Candidate**: for a challenging problem q ∈ qsol, a prefix spre,q is generated from its expert solution sexp,q = (y1, y2, . . . , ym), where m is the total number of tokens in the expert solution. the desired length of prefix, lp, is determined by: lp = ⌊λ · m⌋ where λ is a truncation ratio randomly sampled from...

---

4. Reinforcement learning with dynamic completion for answering multi-hop questions over incomplete knowledge graph

URL: View paper

Brief Assessment

Dynamic Completion[52] focuses on multi-hop question answering over incomplete knowledge graphs using reinforcement learning, which is a different task domain from QuestA's mathematical reasoning with LLMs. No full text was provided for this candidate to assess technical overlap.

---

5. Promed: Shapley information gain guided reinforcement learning for proactive medical llms

URL: View paper

Brief Assessment

Promed[53] focuses on medical LLMs learning to ask diagnostic questions in clinical consultations, using Shapley values to reward information-seeking behavior. This differs fundamentally from QuestA's approach of prepending partial solutions to math problems during RL training to scaffold reasoning tasks.

---

6. Converting Natural Language to Query Languages Using Large Language Models: A Systematic Literature Review

URL: View paper

Brief Assessment

NL to Query[59] focuses on converting natural language to query languages (SQL, GraphQL, etc.) using LLMs, not on reinforcement learning training for reasoning tasks. The candidate addresses a completely different domain (database querying) and does not discuss question augmentation with partial solutions for RL training.

---

7. Enhancing policy gradient for traveling salesman problem with data augmented behavior cloning

URL: View paper

Brief Assessment

TSP Behavior Cloning[54] focuses on data augmentation for the Traveling Salesman Problem using partial solutions in a behavior cloning context, not on question augmentation for general reasoning tasks in LLMs with reinforcement learning.

---

8. Using incomplete and incorrect plans to shape reinforcement learning in long-sequence sparse-reward tasks

URL: View paper

Brief Assessment

Incomplete Plans[51] focuses on using incomplete/incorrect symbolic plans for reward shaping in RL navigation tasks, not on augmenting reasoning problems with partial solutions during training as in QuestA.

---

9. OpenRFT: Adapting Reasoning Foundation Model for Domain-specific Tasks with Reinforcement Fine-Tuning

URL: View paper

Brief Assessment

OpenRFT[57] focuses on question augmentation through rephrasing and option shuffling for domain-specific fine-tuning, not on augmenting questions with partial solutions during RL training. The candidate's data augmentation generates variations while preserving meaning, whereas the original paper prepends partial solution steps to scaffold difficult problems.

Contribution 2: Theoretical analysis of partial-solution augmentation benefits

Description: The paper provides formal theoretical justification showing that augmenting questions with partial solutions (hints) improves RL sample efficiency. The analysis demonstrates that hints enable the model to discover valid trajectories with asymptotically lower sampling budget compared to training without hints.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Guiding Reinforcement Learning with Incomplete System Dynamics

URL: View paper

Brief Assessment

Incomplete Dynamics[67] focuses on incorporating partial system dynamics knowledge into model-free RL for control tasks, not on augmenting training data with partial solutions to improve sample efficiency in reasoning tasks.

---

2. Adaptive Guidance Accelerates Reinforcement Learning of Reasoning Models

URL: View paper

Prior Art Analysis

Adaptive Guidance[61] provides theoretical analysis demonstrating that incorporating hints (partial solutions) into the model's context improves learning efficiency in reinforcement learning for reasoning tasks. The candidate paper presents theoretical frameworks analyzing how natural language guidance enables models to solve problems they could not previously solve, and describes adaptive algorithms that incorporate hints to optimize policy learning. Both papers theoretically justify that augmenting problems with partial solutions improves RL sample efficiency, with the candidate paper providing this analysis prior to the original paper's submission.

Evidence

Evidence 1 - **Rationale**: Both papers provide theoretical justification for why incorporating partial solutions/guidance improves RL training efficiency. The candidate paper derives algorithms based on theoretical insights about guidance improving learning. - **Original**: our theoretical analysis in section 4 explains why partial-solution augmentation accelerates rl training: by decomposing problems into intermediate steps, the method yields denser reward signals and improves sample efficiency, while still driving the model to master the hardest problems. - **Candidate**: we further show that we can significantly improve pass@K rates by leveraging natural language guidance for the model to consider within context while still requiring the model to derive a solution chain from scratch. based of these insights, we derive guide -- a new class of online traini...

Evidence 2 - **Rationale**: Both papers provide formal theoretical results about how hints improve RL learnability and sampling efficiency. The candidate describes adaptive hint incorporation with theoretical grounding. - **Original**: theorem 4.6 (informal upper bound on rl learnability with hint). if we have a hint hq for every question q ∈ q(def. 4.5), then there exists an rl algorithm that can output a policy πθ such that eq∼uniform(q)[pτ∼πθ(·|q)(τ ∈ s(q))] ≥ 0.99 with o(1/δ′ p) sampling budget with high probability. - **Candidate**: guide adaptively incorporates hints into the model's context on problems for which all rollouts were initially incorrect and adjusts the importance sampling ratio for the"off-policy"trajectories in order to optimize the policy for contexts in which the hints are no longer present.

Evidence 3 - **Rationale**: Both papers provide theoretical analysis of learning efficiency when partial solutions are incorporated, demonstrating that the candidate paper established this theoretical framework prior to the original submission. - **Original**: theorem 4.6 provides a theoretical guarantee that the model can reach a high training success rate when partial solution is included. empirically, we observe the model generalizes well both in-distribution and out-of-distribution to hard questions. - **Candidate**: we include careful ablations to analyze guide's components and theoretically analyze guide's learning efficiency.

---

3. StepHint: Multi-level Stepwise Hints Enhance Reinforcement Learning to Reason

URL: View paper

Prior Art Analysis

StepHint[60] provides theoretical analysis demonstrating that partial-solution hints (which they call 'stepwise hints') improve RL sample efficiency by reducing the sampling budget needed to discover valid trajectories. Their Lemma C.1 and Theorem C.4 formally prove that with hints, the sampling budget required is O(1/δ'_p) compared to O(1/δ_p) without hints, where δ'_p = δ_p^(1/2-ϵ). This establishes that hints enable asymptotically lower sampling requirements - essentially the square root of the budget needed without hints. This theoretical framework directly addresses the same problem space as the original paper's contribution, demonstrating prior work on the theoretical benefits of partial-solution augmentation for RL efficiency.

Evidence

Evidence 1 - **Rationale**: Both papers propose augmenting problems with partial solutions (hints) to improve RL training efficiency, establishing that StepHint[60] addresses the same core concept. - **Original**: we introduce quest a, an efficient procedure that controls difficulty by augmenting hard problems with partial solutions. this approach provides a smooth curriculum within rl training and makes high-difficulty tasks more tractable. - **Candidate**: we propose stephint, a novel rlvr algorithm that utilizes multi-level stepwise hints to help models explore the solution space more effectively. stephint generates valid reasoning chains from stronger models and partitions these chains into reasoning steps using our proposed adaptive partitioning me...

---

4. Diffusion-DICE: In-Sample Diffusion Guidance for Offline Reinforcement Learning

URL: View paper

Brief Assessment

Diffusion-DICE[63] focuses on distribution correction estimation methods in offline RL using diffusion models for policy optimization, not on partial-solution hints or augmentation strategies for improving sample efficiency in reasoning tasks.

---

5. ADHint: Adaptive Hints with Difficulty Priors for Reinforcement Learning

URL: View paper

Brief Assessment

ADHint[68] focuses on adaptive hint-ratio scheduling and advantage estimation based on difficulty priors/posteriors, but does not provide formal theoretical analysis showing asymptotically lower sampling budgets. The original paper's theoretical framework (Theorems 4.4, 4.6) with formal definitions of solution sets and capacity sets is distinct from ADHint's empirical approach.

---

6. From Data-Centric to Sample-Centric: Enhancing LLM Reasoning via Progressive Optimization

URL: View paper

Brief Assessment

Sample-Centric[58] does not provide theoretical analysis of why partial solutions improve RL sample efficiency. The candidate focuses on empirical validation and practical implementation details without formal theoretical justification.

---

7. Scalable fragment-based 3d molecular design with reinforcement learning

URL: View paper

Brief Assessment

Fragment-based Design[62] focuses on 3D molecular design using reinforcement learning with fragment placement in physical space, guided by energy-based rewards. It does not address partial solution hints or sample efficiency improvements in general RL contexts.

---

8. Sample Efficient Reinforcement Learning with Partial Dynamics Knowledge

URL: View paper

Brief Assessment

Partial Dynamics[65] addresses partial knowledge of system dynamics (f function) in RL, not partial solutions as hints. The theoretical framework analyzes how knowing f improves sample efficiency, which is orthogonal to QUESTA's analysis of hint-based augmentation for reasoning tasks.

---

9. DRlinker: deep reinforcement learning for optimization in fragment linking design

URL: View paper

Brief Assessment

DRlinker[64] focuses on fragment linking in drug design using RL to optimize molecular properties, not on partial-solution hints for improving RL sample efficiency in reasoning tasks. The theoretical frameworks address entirely different domains (chemistry vs. reasoning).

---

10. Integrating reaction schemes, reagent databases, and virtual libraries into fragment-based design by reinforcement learning

URL: View paper

Brief Assessment

Fragment Linking[66] focuses on fragment-based drug design using reinforcement learning for molecular optimization, not on partial-solution hints for improving RL sample efficiency in reasoning tasks. The domains and technical approaches are fundamentally different.

Contribution 3: State-of-the-art results for 1.5B-parameter models on math benchmarks

Description: The authors demonstrate that applying QuestA to small-scale models (1.5B parameters) achieves new state-of-the-art performance on challenging mathematical reasoning benchmarks, substantially outperforming existing models of similar size and even matching or exceeding much larger 32B-parameter models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking

URL: View paper

Prior Art Analysis

rStar-Math[69] demonstrates that small language models (1.5B-3.8B parameters) can achieve state-of-the-art mathematical reasoning performance that rivals or surpasses OpenAI o1 models. The candidate paper shows qwen2.5-math-1.5b achieving 88.6% on MATH and 46.7% on AIME 2024, while phi3-mini-3.8b reaches 86.4% on MATH and 43.3% on AIME 2024. These results directly refute the novelty claim that QuestA was the first to achieve state-of-the-art performance with 1.5B-parameter models, as rStar-Math demonstrates comparable or superior performance with models of similar or smaller size through a different methodology (MCTS-based deep thinking vs. question augmentation with partial solutions).

Evidence

Evidence 1 - **Rationale**: Both papers claim state-of-the-art results for small parameter models on challenging math benchmarks. rStar-Math demonstrates that models as small as 1.5B-3.8B parameters can achieve performance comparable to or exceeding the original paper's 1.5B model results. - **Original**: we achieve new state-of-the-art results on math benchmarks using 1.5b-parameter models: 72.50% (+10.73%) on aime24, 62.29% (+12.79%) on aime25, and 41.67% (+10.11%) on hmmt25. - **Candidate**: on the math benchmark, it improves qwen2.5-math-7b from 58.8% to 90.0% and phi3-mini-3.8b from 41.4% to 86.4%, surpassing o1-preview by +4.5% and +0.9%. on the usa math olympiad (aime), rstar-math solves an average of 53.3% (8/15) of problems, ranking among the top 20% the brightest high school math...

Evidence 2 - **Rationale**: Both papers explicitly claim to achieve 'state-of-the-art' results for small language models on math benchmarks, with rStar-Math demonstrating this achievement through a different technical approach (MCTS-based reasoning) rather than question augmentation. - **Original**: quest a is a data augmentation method that injects partial solutions to effectively scaffold rl training on hard reasoning problems. we construct 26k high-quality augmented prompts from challenging instances in openr1 (open-r1 team, 2025), and fine-tune models using 32kcontext-length rl. when applie... - **Candidate**: rstar-math boosts slms' math reasoning to state-of-the-art levels. on the math benchmark, it improves qwen2.5-math-7b from 58.8% to 90.0% and phi3-mini-3.8b from 41.4% to 86.4%, surpassing o1-preview by +4.5% and +0.9%.

Evidence 3 - **Rationale**: The table shows rStar-Math achieving 88.6% on MATH with a 1.5B model, demonstrating state-of-the-art performance for small models that directly challenges the original paper's novelty claim of being first to achieve such results with 1.5B parameters. - **Original**: our quest a-nemotron-1.5b achieves state-of-the-art performance among 1.5b models and, notably, matches or even exceeds the performance of deepseek-r1-distill-32b across several benchmarks, despite being over 20x smaller in parameter count. - **Candidate**: rstar-math (qwen-1.5b) rstar-math (phi3-mini) openai o1-preview openai o1-mini qwq 32b-previewgpt-4odeepseek-v3 math 90.0 88.6 86.4 85.5 90.0 90.6 76.6 90.2 aime 2024 53.3 46.7 43.3 44.6 56.7 50.0 9.3 39.2

Evidence 4 - **Rationale**: Both papers demonstrate substantial improvements on challenging math benchmarks using small models, with rStar-Math showing comparable or superior absolute performance levels despite using a different methodology. - **Original**: table 1 reports results on challenging math benchmarks. quest a yields substantial gains for nemotron-1.5b, achieving an average improvement of 10% over its baseline and a particularly strong +13% on aime25. these improvements are consistent across all benchmarks, highlighting the effectiveness of o... - **Candidate**: through 4 rounds of self-evolution with millions of synthesized solutions for 747k math problems, rstar-math boosts slms' math reasoning to state-of-the-art levels. on the math benchmark, it improves qwen2.5-math-7b from 58.8% to 90.0% and phi3-mini-3.8b from 41.4% to 86.4%

---

2. Orca 2: Teaching Small Language Models How to Reason

URL: View paper

Brief Assessment

Orca 2[75] focuses on teaching small models (7B-13B parameters) various reasoning strategies through explanation tuning and cautious reasoning, not specifically on 1.5B-parameter models achieving state-of-the-art mathematical reasoning performance through question augmentation with partial solutions as in the original paper.

3. Jiuzhang3. 0: Efficiently improving mathematical reasoning by training small data synthesis models

URL: View paper

Brief Assessment

Jiuzhang[76] focuses on training small data synthesis models (7B) to generate pre-training data, achieving SOTA with 7B/8B models. The candidate's approach differs fundamentally: it uses RL with question augmentation (partial solutions) rather than synthetic data generation for pre-training.

4. Specializing smaller language models towards multi-step reasoning

URL: View paper

Brief Assessment

Specializing Small LLMs[70] focuses on distilling chain-of-thought reasoning from GPT-3.5 to T5 models (≤11B parameters) for multi-step math reasoning, achieving improvements through model specialization. However, this work predates the ORIGINAL paper and does not address the specific QuestA method, reinforcement learning with question augmentation, or the particular benchmarks (AIME24, AIME25, HMMT25) where the ORIGINAL paper claims state-of-the-art results for 1.5B models.

5. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution

URL: View paper

Brief Assessment

SWE-RL[37] focuses on software engineering tasks (GitHub issue resolution) rather than mathematical reasoning benchmarks. The candidate achieves 41.0% on SWE-Bench Verified but does not report results on math benchmarks like AIME or HMMT that are central to the original paper's contribution.

6. LLM performance on mathematical reasoning in Catalan language

URL: View paper

Brief Assessment

Catalan Math[74] evaluates existing LLMs on Catalan-language mathematical problems from the Kangaroo Mathematics Competition, focusing on language-specific performance rather than advancing state-of-the-art results for small-parameter models on standard mathematical reasoning benchmarks.

7. Tiny-R1V: Lightweight Multimodal Unified Reasoning Model via Model Merging

URL: View paper

Brief Assessment

Tiny-R1V[77] focuses on a 3B multimodal model achieving efficiency through model merging and length-informed optimization, not specifically on 1.5B parameter models achieving state-of-the-art math performance through question augmentation as in the original paper.

8. Scalediff: Scaling difficult problems for advanced mathematical reasoning

URL: View paper

Brief Assessment

Scalediff[72] focuses on a 7B-parameter model (scalediff-7b) achieving 65.9% average accuracy, not 1.5B models. The paper does not demonstrate state-of-the-art results for 1.5B-parameter models specifically.

9. Small models struggle to learn from strong reasoners

URL: View paper

Brief Assessment

Small Models Struggle[73] focuses on the learnability gap in small models when distilling from strong reasoners, not on achieving state-of-the-art performance through question augmentation methods like QuestA.

10. Chain of draft: Thinking faster by writing less

URL: View paper

Brief Assessment

Chain of Draft[71] focuses on reducing token usage and latency in reasoning through concise intermediate steps, not on achieving state-of-the-art performance for small-scale models on mathematical benchmarks through reinforcement learning methods.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] QuestA: Expanding Reasoning Capacity in LLMs via Question Augmentation View paper
- [1] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? View paper
- [2] Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models View paper
- [3] Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning View paper
- [4] Learning to reason with search for llms via reinforcement learning View paper
- [5] Reinforcement learning for reasoning in large language models with one training example View paper
- [6] Advancing reasoning in large language models: Promising methods and approaches View paper
- [7] Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning View paper
- [8] Towards large reasoning models: A survey of reinforced reasoning with large language models View paper
- [9] Teaching Large Language Models to Reason with Reinforcement Learning View paper
- [10] Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning View paper
- [11] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [12] Rethinking Reasoning Quality in Large Language Models through Enhanced Chain-of-Thought via RL View paper
- [13] Fine-tuning large vision-language models as decision-making agents via reinforcement learning View paper
- [14] Effective Reinforcement Learning for Reasoning in Language Models View paper
- [15] A survey of reinforcement learning for large reasoning models View paper

- [16] Demystifying long chain-of-thought reasoning in llms View paper
- [17] Grounding Large Language Models in Interactive Environments with Online Reinforcement Learning View paper
- [18] WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct View paper
- [19] Dapo: An open-source llm reinforcement learning system at scale View paper
- [20] A survey on large language models for mathematical reasoning View paper
- [21] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't View paper
- [22] Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning View paper
- [23] Enhance Reasoning for Large Language Models in the Game Werewolf View paper
- [24] Offline reinforcement learning for llm multi-step reasoning View paper
- [25] Advancing language model reasoning through reinforcement learning and inference scaling View paper
- [26] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms View paper
- [27] A technical survey of reinforcement learning techniques for large language models View paper
- [28] Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search View paper
- [29] DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning View paper
- [30] Multi-step reasoning with large language models, a survey View paper
- [31] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning View paper
- [32] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning View paper
- [33] Training Large Language Models for Reasoning through Reverse Curriculum Reinforcement Learning View paper
- [34] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router View paper
- [35] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle View paper
- [36] Interleaved Reasoning for Large Language Models via Reinforcement Learning View paper
- [37] Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution View paper
- [38] Reinforced Latent Reasoning for LLM-based Recommendation View paper
- [39] Reasoning with Exploration: An Entropy Perspective View paper
- [40] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models View paper
- [41] Guiding Pretraining in Reinforcement Learning with Large Language Models View paper
- [42] Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning View paper
- [43] A Collaborative Reasoning Framework Powered by Reinforcement Learning and Large Language Models for Complex Questions Answering over Knowledge Graph View paper
- [44] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning View paper
- [45] Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment View paper
- [46] Eliciting Chain-of-Thought Reasoning for Time Series Analysis using Reinforcement Learning View paper
- [47] An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents View paper
- [48] Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning View paper
- [49] Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks View paper
- [50] Rewarding Progress: Scaling Automated Process Verifiers for LLM Reasoning View paper
- [51] Using incomplete and incorrect plans to shape reinforcement learning in long-sequence sparse-reward tasks View paper
- [52] Reinforcement learning with dynamic completion for answering multi-hop questions over incomplete knowledge graph View paper
- [53] Promed: Shapley information gain guided reinforcement learning for proactive medical llms View paper
- [54] Enhancing policy gradient for traveling salesman problem with data augmented behavior cloning View paper
- [55] Mixture of Autoencoder Experts Guidance using Unlabeled and Incomplete Data for Exploration in Reinforcement Learning View paper
- [56] Generative question refinement with deep reinforcement learning in retrieval-based QA system View paper
- [57] OpenRFT: Adapting Reasoning Foundation Model for Domain-specific Tasks with Reinforcement Fine-Tuning View paper
- [58] From Data-Centric to Sample-Centric: Enhancing LLM Reasoning via Progressive Optimization View paper
- [59] Converting Natural Language to Query Languages Using Large Language Models: A Systematic Literature Review View paper
- [60] StepHint: Multi-level Stepwise Hints Enhance Reinforcement Learning to Reason View paper
- [61] Adaptive Guidance Accelerates Reinforcement Learning of Reasoning Models View paper
- [62] Scalable fragment-based 3d molecular design with reinforcement learning View paper
- [63] Diffusion-DICE: In-Sample Diffusion Guidance for Offline Reinforcement Learning View paper
- [64] DRlinker: deep reinforcement learning for optimization in fragment linking design View paper
- [65] Sample Efficient Reinforcement Learning with Partial Dynamics Knowledge View paper
- [66] Integrating reaction schemes, reagent databases, and virtual libraries into fragment-based design by reinforcement learning View paper
- [67] Guiding Reinforcement Learning with Incomplete System Dynamics View paper
- [68] ADHint: Adaptive Hints with Difficulty Priors for Reinforcement Learning View paper
- [69] rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking View paper
- [70] Specializing smaller language models towards multi-step reasoning View paper
- [71] Chain of draft: Thinking faster by writing less View paper
- [72] Scalediff: Scaling difficult problems for advanced mathematical reasoning View paper
- [73] Small models struggle to learn from strong reasoners View paper
- [74] LLM performance on mathematical reasoning in Catalan language View paper
- [75] Orca 2: Teaching Small Language Models How to Reason View paper
- [76] Jiuzhang3. 0: Efficiently improving mathematical reasoning by training small data synthesis models View paper
- [77] Tiny-R1V: Lightweight Multimodal Unified Reasoning Model via Model Merging View paper