Novelty Assessment Report

Paper: LoongRL: Reinforcement Learning for Advanced Reasoning over Long Contexts

PDF URL: https://openreview.net/pdf?id=o29E01Q6bv

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Reasoning over long contexts is essential for large language models. While reinforcement learning (RL) enhances short-context reasoning by inducing "Aha" moments in chain-of-thought, the advanced thinking patterns required for long-context reasoning remain largely unexplored, and high-difficulty RL data are scarce. In this paper, we introduce LoongRL, a data-driven RL method for advanced long-context reasoning. Central to LoongRL is KeyChain, a synthesis approach that transforms short multi-hop QA into high-difficulty long-context tasks by inserting UUID chains that hide the true question among large collections of distracting documents. Solving these tasks requires the model to trace the correct chain step-by-step, identify the true question, retrieve relevant facts and reason over them to answer correctly. RL training on KeyChain data induces an emergent plan–retrieve–reason–recheck reasoning pattern that generalizes far beyond training length. Models trained at 16K effectively solve 128K tasks without prohibitive full-length RL rollout costs. On Qwen2.5-7B and 14B, LoongRL substantially improves long-context multi-hop QA accuracy by +23.5% and +21.1% absolute gains. The resulting LoongRL-14B reaches a score of 74.2, rivaling much larger frontier models such as o3-mini (74.5) and DeepSeek-R1 (74.9). It also improves long-context retrieval, passes all 128K needle-in-a-haystack stress tests, and preserves short-context reasoning capabilities.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Reinforcement Learning for Long-Context Reasoning in Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

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

- **RL Training Methodologies for Reasoning Enhancement**
- **Long-Context Reasoning and Memory Management**
- **Inference-Time Computation and Scaling**
- **Application Domains and Task-Specific Adaptations**
- **Post-Training and Model Optimization**

Complete Taxonomy Tree

- Reinforcement Learning for Long-Context Reasoning in Language Models Survey Taxonomy
- RL Training Methodologies for Reasoning Enhancement
  - General RL Frameworks and Algorithm Design (6 papers)
  - [1] A survey of reinforcement learning for large reasoning models (Zhang Kai-Yan, 2025) View paper
  - [10] Effective Reinforcement Learning for Reasoning in Language Models (Li Shuo, 2025) View paper
  - [11] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle (Liu Ke-liang, 2025) View paper
  - [12] Advancing Language Model Reasoning through Reinforcement Learning and Inference Scaling (Hou Zhenyu, 2025) View paper
  - [34] The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models (Cui, 2025) View paper
  - [46] Reinforcement learning: Advanced techniques for llm behavior optimization (Hariharan, 2025) View paper
  - Process-Level Supervision and Reward Modeling (3 papers)
  - [6] Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents (Zhang Zijing, 2025) View paper
  - [15] Think-RM: Enabling Long-Horizon Reasoning in Generative Reward Models (Hong, 2025) View paper
  - [43] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning (Ning, 2025) View paper
  - Multi-Turn and Multi-Agent RL Training (5 papers)
  - [19] Context-lite multi-turn reinforcement learning for LLM agents (W Chen, 2025) View paper
  - [30] Agentic Reinforced Policy Optimization (Dong, 2025) View paper
  - [33] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning (Xi, 2025) View paper
  - [40] DeepPlanner: Scaling Planning Capability for Deep Research Agents via Advantage Shaping (Fan Wei, 2025) View paper
  - [50] PilotRL: Training Language Model Agents via Global Planning-Guided Progressive Reinforcement Learning (Chen Chong, 2025) View paper
  - Self-Play and Exploration Mechanisms (2 papers)
  - [2] Spell: Self-play reinforcement learning for evolving long-context language models (Yang Ziyi, 2025) View paper
  - [32] Satori: Reinforcement Learning with Chain-of-Action-Thought Enhances LLM Reasoning via Autoregressive Search (Shen, 2025) View paper
- Long-Context Reasoning and Memory Management
  - RL-Based Long-Context Training ★ (3 papers)

  - [0] LoongRL: Reinforcement Learning for Advanced Reasoning over Long Contexts (Anon et al., 2026) View paper
  - [25] QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning (Wan, 2025) View paper
  - [49] QwenLong-L1.5: Post-Training Recipe for Long-Context Reasoning and Memory Management (Weizhou Shen, 2025) View paper
  - Memory Architectures and External Memory Banks (4 papers)
  - [22] Memory-T1: Reinforcement Learning for Temporal Reasoning in Multi-session Agents (Yiming Du, 2025) View paper
  - [23] MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent (Yu Hongli, 2025) View paper
  - [38] Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning (Yang Xiufeng, 2025) View paper
  - [48] Look back to reason forward: Revisitable memory for long-context llm agents (Chen, 2025) View paper
  - Context Compression and Summarization (1 papers)
  - [44] Scaling LLM Multi-turn RL with End-to-end Summarization-based Context Management (Lu Miao, 2025) View paper
- Inference-Time Computation and Scaling
  - Test-Time Scaling and Search Strategies (2 papers)
  - [17] A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law (Qianjun Pan, 2025) View paper
  - [47] R-Search: Empowering LLM Reasoning with Search via Multi-Reward Reinforcement Learning (Zhao Qing-fei, 2025) View paper
  - Parallel Reasoning and Coordination (1 papers)
  - [4] Learning Adaptive Parallel Reasoning with Language Models (Pan Jiayi, 2025) View paper
  - In-Context Learning and Reinforcement (3 papers)
  - [16] Reward Is Enough: LLMs Are In-Context Reinforcement Learners (Kefan Song, 2025) View paper
  - [20] Llms are in-context reinforcement learners (Giovanni Monea, 2024) View paper
  - [35] Beyond Numeric Rewards: In-Context Dueling Bandits with LLM Agents (Fanzeng Xia, 2024) View paper
- Application Domains and Task-Specific Adaptations
  - Tool Integration and External Knowledge Access (4 papers)
  - [7] Reinforcement learning foundations for deep research systems: A survey (Li Wen-Jun, 2025) View paper
  - [8] Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning (Jin Bo-wen, 2025) View paper
  - [14] WebExplorer: Explore and Evolve for Training Long-Horizon Web Agents (Liu Jun-teng, 2025) View paper
  - [29] Reinforcement Fine-Tuning for Reasoning towards Multi-Step Multi-Source Search in Large Language Models (Shi Wentao, 2025) View paper
  - Interactive Environments and Long-Horizon Agents (2 papers)
  - [9] Reinforcement Learning for Long-Horizon Interactive LLM Agents (Chen Kevin, 2025) View paper
  - [45] LLM-Guided Reinforcement Learning for Interactive Environments (Fuxue Yang, 2025) View paper
  - Mathematical and Formal Reasoning (2 papers)
  - [21] JT-Math: A Multi-Stage Framework for Advanced Mathematical Reasoning in Large Language Models (Hao Yi-fan, 2025) View paper
  - [41] DeepTheorem: Advancing LLM Reasoning for Theorem Proving Through Natural Language and Reinforcement Learning (Zhang Zi-yin, 2025) View paper
  - Multi-Modal and Vision-Language Reasoning (3 papers)
  - [27] SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning (Zuo Yuxin, 2025) View paper
  - [37] Hierarchical DLO Routing with Reinforcement Learning and In-Context Vision-language Models (Li MingEn, 2025) View paper
  - [39] GFlowVLM: Enhancing Multi-step Reasoning in Vision-Language Models with Generative Flow Networks (Kang, 2025) View paper
  - Specialized Task Domains (4 papers)
  - [26] R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning (He Minggui, 2025) View paper
  - [28] Leveraging Large Language Model for Intelligent Log Processing and Autonomous Debugging in Cloud AI Platforms (Ji, 2025) View paper
  - [31] E3-Rewrite: Learning to Rewrite SQL for Executability, Equivalence,and Efficiency (Xu Dongjie, 2025) View paper
  - [36] Contextual integrity in llms via reasoning and reinforcement learning (Inan, 2025) View paper
- Post-Training and Model Optimization
  - Multi-Stage Post-Training Frameworks (2 papers)
  - [3] Llm post-training: A deep dive into reasoning large language models (Kumar, 2025) View paper
  - [5] Kimi k1.5: Scaling Reinforcement Learning with LLMs (Kimi Team, 2025) View paper
  - Structured Data Processing and Representation Learning (1 papers)
  - [24] Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models (Lee YoungHun, 2024) View paper
  - Resource-Constrained and Edge Deployment (2 papers)
  - [18] Serving Long-Context LLMs at the Mobile Edge: Test-Time Reinforcement Learning-based Model Caching and Inference Offloading (Xu, 2025) View paper
  - [42] Cache-Efficient Posterior Sampling for Reinforcement Learning with LLM-Derived Priors Across Discrete and Continuous Domains (Ibne Farabi Shihab, 2025) View paper
  - Multi-Agent Collaboration and Theory of Mind (1 papers)
  - [13] Theory of Mind for Multi-Agent Collaboration via Large Language Models (Huao Li, 2023) View paper

Narrative

Core task: reinforcement learning for long-context reasoning in language models. The field has organized itself around several complementary branches that address different facets of this challenge. RL Training Methodologies for Reasoning Enhancement focuses on algorithmic innovations—policy gradient techniques, reward shaping, and self-play mechanisms—that enable models to learn complex reasoning behaviors, as seen in works like Spell Self-play[2] and Effective RL Reasoning[10]. Long-Context Reasoning and Memory Management tackles the architectural and data-handling side, exploring how models can maintain coherent reasoning over extended sequences through memory mechanisms and context compression strategies, exemplified by Kimi[5] and QwenLong-L1[25]. Inference-Time Computation and Scaling examines how to allocate computational resources during test time—via search, iterative refinement, or adaptive depth—to improve reasoning quality without retraining, while Application Domains and Task-Specific Adaptations and Post-Training and Model Optimization address deployment contexts and fine-tuning recipes that bridge research prototypes and production systems.

Within this landscape, a particularly active line of work centers on integrating RL directly into long-context training pipelines, balancing the need for extended memory with the sample efficiency and stability challenges of reinforcement learning. LoongRL[0] sits squarely in this cluster, emphasizing RL-based training specifically designed for long-context scenarios, closely aligned with QwenLong Recipe[49] and QwenLong-L1[25], which also explore how to scale context windows while maintaining reasoning fidelity through RL-driven optimization. In contrast, broader surveys like RL Large Reasoning Survey[1] and LLM Post-training Deep Dive[3] provide overarching perspectives on reasoning enhancement and post-training strategies, situating long-context RL as one specialized direction among many. The central tension across these branches remains how to efficiently train models that reason deeply over long horizons without prohibitive computational costs or unstable learning dynamics, a question that LoongRL[0] addresses by focusing on RL techniques tailored to extended context lengths.

Related Works in Same Category

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

1. QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning

Authors: Wan, Fanqi, Shen Weizhou, Fanqi Wan, Weizhou Shen, et al. (22 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent large reasoning models (LRMs) have demonstrated strong reasoning capabilities through reinforcement learning (RL). These improvements have primarily been observed within the short-context reasoning tasks. In contrast, extending LRMs to effectively process and reason on long-context inputs via RL remains a critical unsolved challenge. To bridge this gap, we first formalize the paradigm of long-context reasoning RL, and identify key challenges in suboptimal training efficiency and unstable ...

Relationship Analysis

Both papers belong to the RL-Based Long-Context Training category, focusing on applying reinforcement learning to train models for long-context reasoning tasks. They overlap in their core approach of using RL to improve long-context multi-hop question answering and both demonstrate generalization from shorter training contexts to longer inference contexts. The key differences are that LoongRL introduces the KeyChain data synthesis method with UUID chains to create high-difficulty training data and trains at 16K context length, while QwenLong-L1 emphasizes progressive context scaling with curriculum-guided phased RL and trains on sequences up to 60K tokens, using different strategies for stabilizing the RL optimization process.

2. QwenLong-L1.5: Post-Training Recipe for Long-Context Reasoning and Memory Management

Authors: Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

We introduce QwenLong-L1.5, a model that achieves superior long-context reasoning capabilities through systematic post-training innovations. The key technical breakthroughs of QwenLong-L1.5 are as follows: (1) Long-Context Data Synthesis Pipeline: We develop a systematic synthesis framework that generates challenging reasoning tasks requiring multi-hop grounding over globally distributed evidence. By deconstructing documents into atomic facts and their underlying relationships, and then programm...

Relationship Analysis

Both papers belong to the RL-Based Long-Context Training category, focusing on reinforcement learning methods to enhance long-context reasoning in language models. They share overlapping approaches in using RL for training on extended sequences, synthetic data construction for long-context tasks, and addressing the challenge of generalizing from shorter training contexts to longer inference contexts. The key differences are that LoongRL introduces the KeyChain data synthesis method with UUID chains to induce plan-retrieve-reason-recheck patterns and trains at 16K to generalize to 128K, while QwenLong-L1.5 emphasizes a systematic fact-based data synthesis pipeline, proposes Adaptive Entropy-Controlled Policy Optimization (AEPO) for training stability, and incorporates a memory-augmented architecture for ultra-long contexts exceeding 4M tokens.

Contributions Analysis

Overall novelty summary. The paper introduces LoongRL, a data-driven RL method for long-context reasoning, alongside KeyChain, a synthesis approach that transforms short multi-hop QA into high-difficulty long-context tasks using UUID chains. Within the taxonomy, this work resides in the 'RL-Based Long-Context Training' leaf under 'Long-Context Reasoning and Memory Management'. This leaf contains only three papers total, including the original work, indicating a relatively sparse research direction. The sibling papers (QwenLong Recipe and QwenLong-L1) also focus on RL-driven optimization for extended context windows, suggesting this is an emerging but not yet crowded subfield.

The taxonomy reveals that neighboring leaves address complementary challenges: 'Memory Architectures and External Memory Banks' explores external memory systems for long-horizon reasoning, while 'Context Compression and Summarization' focuses on reducing context length. The broader parent branch 'Long-Context Reasoning and Memory Management' sits alongside 'RL Training Methodologies for Reasoning Enhancement', which houses general-purpose RL frameworks and process-level supervision methods. LoongRL bridges these areas by applying RL specifically to long-context scenarios, diverging from general short-context RL methods and memory-based architectures that do not emphasize RL training.

Among 30 candidates examined, none clearly refute the three core contributions: the LoongRL method (10 candidates, 0 refutable), the KeyChain synthesis approach (10 candidates, 0 refutable), and the two-way substring exact match verifier (10 candidates, 0 refutable). This suggests that within the limited search scope, the specific combination of data-driven RL for long-context reasoning, UUID-chain-based task synthesis, and the proposed verification mechanism appears novel. However, the search scale is modest, and the analysis does not claim exhaustive coverage of all potentially relevant prior work.

Given the sparse taxonomy leaf and the absence of refuting candidates among the 30 examined, the work appears to occupy a relatively unexplored niche within long-context RL training. The limited search scope means this assessment is provisional; a broader literature review might uncover additional overlapping methods. Nonetheless, the combination of RL-driven training, synthetic task generation via UUID chains, and emergent reasoning patterns at extended lengths represents a distinctive approach within the current field structure.

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

Contribution 1: LoongRL: data-driven RL method for advanced long-context reasoning

Description: The authors propose LoongRL, a reinforcement learning approach that enables models to acquire effective thinking patterns for long-context reasoning tasks. The method trains models to develop emergent plan-retrieve-reason-recheck reasoning patterns that generalize beyond training length.

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. Large language models are learnable planners for long-term recommendation

URL: View paper

Brief Assessment

LLM Planners Recommendation[57] focuses on reinforcement learning for long-term recommendation systems with planning for user engagement, not long-context reasoning in language models. The technical domains are fundamentally different.

2. Amago: Scalable in-context reinforcement learning for adaptive agents

URL: View paper

Brief Assessment

AMAGO[56] focuses on in-context RL for meta-learning and goal-conditioned problems in procedurally generated environments, not on long-context reasoning for language models or multi-hop QA tasks.

3. Spell: Self-play reinforcement learning for evolving long-context language models

URL: View paper

Brief Assessment

Spell Self-play[2] focuses on a multi-role self-play framework (questioner-responder-verifier) for label-free optimization, whereas LoongRL centers on keychain data synthesis and emergent plan-retrieve-reason-recheck patterns. The technical approaches and data construction methods differ fundamentally.

4. Retrieval-augmented hierarchical in-context reinforcement learning and hindsight modular reflections for task planning with llms

URL: View paper

Brief Assessment

Hierarchical In-Context RL[54] focuses on hierarchical task decomposition for embodied decision-making in robotics and web navigation environments (ALFWorld, WebShop, HotpotQA), not on training language models for long-context reasoning through reinforcement learning. The candidate addresses multi-step task planning with LLMs as controllers, while the original develops RL methods to train models to acquire reasoning patterns for processing extended text contexts.

5. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models

URL: View paper

Brief Assessment

ProRL[51] focuses on prolonged RL training to expand reasoning boundaries in general reasoning tasks, not specifically on long-context reasoning with retrieval from extended documents. The candidate does not address the keychain data construction method or plan-retrieve-reason-recheck patterns for long-context tasks.

6. Chain of agents: Large language models collaborating on long-context tasks

URL: View paper

Brief Assessment

Chain of Agents[52] focuses on multi-agent collaboration for long-context tasks through sequential communication between agents, not on reinforcement learning methods for training models to develop reasoning patterns. The candidate addresses long-context processing through agent coordination rather than RL-based training approaches.

7. Large Language Models Post-training: Surveying Techniques from Alignment to Reasoning

URL: View paper

Brief Assessment

Post-training Survey[58] discusses general RL paradigms for alignment but does not present a specific method for long-context reasoning with emergent plan-retrieve-reason-recheck patterns or keychain data construction.

8. The pokeagent challenge: Competitive and long-context learning at scale

URL: View paper

Brief Assessment

PokeAgent Challenge[53] focuses on competitive game-playing and RPG speedrunning in Pokémon environments, not on general long-context reasoning methods for language models. The candidate addresses different tasks (game AI, opponent modeling, exploration) rather than the reasoning patterns for document-based QA that LoongRL targets.

9. Robohorizon: An llm-assisted multi-view world model for long-horizon robotic manipulation

URL: View paper

Brief Assessment

RoboHorizon[55] focuses on visual model-based RL for robotic manipulation tasks with multi-view world models, not on language model reasoning over long text contexts. The domains and technical approaches are fundamentally different.

10. Kimi k1.5: Scaling Reinforcement Learning with LLMs

URL: View paper

Brief Assessment

Kimi[5] focuses on general RL training for multi-modal LLMs across diverse reasoning tasks (math, code, vision), while LoongRL specifically targets long-context reasoning with a novel keychain data synthesis approach. The technical approaches and problem scopes differ substantially.

Contribution 2: KeyChain: synthesis approach for high-difficulty long-context tasks

Description: The authors introduce KeyChain, a data synthesis method that converts short multi-hop question-answering tasks into challenging long-context problems by inserting UUID chains that hide the true question among distracting documents, requiring models to trace chains step-by-step.

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. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl

URL: View paper

Brief Assessment

DeepDive[75] focuses on synthesizing complex questions from knowledge graphs for web browsing agents, not on transforming short multi-hop QA into long-context tasks using UUID chains. The technical approaches and application domains differ fundamentally.

2. Generalizing from short to long: Effective data synthesis for long-context instruction tuning

URL: View paper

Brief Assessment

Short to Long[76] focuses on context synthesis by generating extended background contexts for existing instruction-answer pairs, while KeyChain creates high-difficulty tasks by inserting UUID chains that require step-by-step tracing. These are fundamentally different synthesis approaches with different objectives and mechanisms.

3. Plan-and-Act: Improving Planning of Agents for Long-Horizon Tasks

URL: View paper

Brief Assessment

Plan-and-Act[70] focuses on web navigation tasks using a planner-executor architecture with synthetic plan annotations for trajectories. This is fundamentally different from KeyChain's approach of transforming multi-hop QA into long-context reasoning tasks using UUID chains and distractor documents.

4. Wildlong: Synthesizing realistic long-context instruction data at scale

URL: View paper

Brief Assessment

WildLong[73] focuses on extracting meta-information from real user queries and graph-based co-occurrence modeling for diverse instruction synthesis, not on UUID chain insertion for multi-hop QA transformation. The technical approaches are fundamentally different.

5. WebExplorer: Explore and Evolve for Training Long-Horizon Web Agents

URL: View paper

Brief Assessment

WebExplorer[14] focuses on web navigation and information-seeking tasks through model-based exploration and query evolution, not on transforming short multi-hop QA into long-context tasks via UUID chain insertion as KeyChain does.

6. Multi-Document Grounded Multi-Turn Synthetic Dialog Generation

URL: View paper

Brief Assessment

Multi-Document Dialog[77] focuses on multi-turn dialog generation with document grounding and retrieval updates, not on transforming short multi-hop QA into long-context tasks via UUID chain insertion as KeyChain does.

7. Generating Multi-turn Clarification for Web Information Seeking

URL: View paper

Brief Assessment

Multi-turn Clarification[74] focuses on generating clarifying questions for web search to handle ambiguous user intents, not on transforming short multi-hop QA into long-context reasoning tasks through data synthesis methods like UUID chains.

8. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices

URL: View paper

Brief Assessment

Long Context Multihop[71] focuses on multi-agent interactive generation of multi-hop questions from documents, not on UUID chain insertion to hide questions among distractors. The candidate's approach involves merging single-hop questions into multi-hop queries, whereas KeyChain transforms tasks by inserting UUID chains that require step-by-step tracing to recover hidden questions.

9. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks

URL: View paper

Brief Assessment

LongBench v2[69] focuses on benchmark construction with human-annotated multi-choice questions requiring deep understanding across diverse realistic tasks. It does not propose a data synthesis method for transforming short multi-hop QA into long-context tasks via UUID chains.

10. Odysseybench: Evaluating llm agents on long-horizon complex office application workflows

URL: View paper

Brief Assessment

OdysseyBench[72] focuses on evaluating LLM agents on long-horizon workflows in office applications (Word, Excel, PDF, etc.), not on transforming short multi-hop QA into long-context reasoning tasks through UUID chain insertion as KeyChain does.

Contribution 3: Two-way substring exact match verifier for RL training

Description: The authors design a rule-based reward verification method that checks whether the extracted answer contains the ground truth as a substring or vice versa, enabling reliable reinforcement learning training on general question-answering tasks without requiring LLM-based judgment.

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. Sparks of tabular reasoning via text2sql reinforcement learning

URL: View paper

Brief Assessment

Tabular Reasoning[61] focuses on text-to-SQL tasks with symbolic reasoning and SQL syntax validation, not general question-answering with substring-based answer verification for RL training.

2. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning

URL: View paper

Brief Assessment

Logic-RL[60] uses a 'stringent format reward function' for logic puzzles with controllable verification, while the original paper's two-way substring exact match specifically addresses free-form general QA answers where multiple valid phrasings exist.

3. DialogueReason: Rule-Based RL Sparks Dialogue Reasoning in LLMs

URL: View paper

Brief Assessment

DialogueReason[67] focuses on dialogue-based reasoning with rule-based rewards for PPO training, but does not describe a two-way substring exact match verifier for answer verification in general QA tasks.

4. SATURN: SAT-based Reinforcement Learning to Unleash LLMs Reasoning

URL: View paper

Brief Assessment

SATURN[63] focuses on SAT-based RL tasks with rule-based verification for SAT problems, not on general question-answering with substring matching for free-form answers. The verification mechanisms serve different problem domains and answer formats.

5. Prompt-Based Length Controlled Generation with Reinforcement Learning

URL: View paper

Brief Assessment

Length Controlled Generation[68] uses rule-based rewards for length control in summarization tasks, not for answer verification in question-answering. The technical focus and application domain differ fundamentally from the original paper's two-way substring exact match verifier for QA tasks.

6. VerIF: Verification Engineering for Reinforcement Learning in Instruction Following

URL: View paper

Brief Assessment

VerIF[59] focuses on instruction-following tasks using a combination of rule-based code verification and LLM-based verification, whereas the original paper's two-way substring exact match is specifically designed for general question-answering tasks with free-form answers in long-context reasoning scenarios.

7. DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding

URL: View paper

Brief Assessment

DocThinker[64] uses rule-based rewards for RL in document understanding but does not describe a two-way substring exact match verifier. The candidate focuses on multi-objective rule-based rewards for explainability in multimodal document tasks, not on answer verification methods for general QA.

8. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning

URL: View paper

Brief Assessment

MM-Eureka[65] focuses on multimodal mathematical reasoning with rule-based RL, but does not describe the specific answer verification mechanism used. The candidate's brief abstract mentions 'rule-based reinforcement learning' but provides no details about substring matching or answer verification methods that would refute the original paper's novelty claim.

9. WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning

URL: View paper

Brief Assessment

WeThink[66] focuses on multimodal visual-language reasoning with a hybrid reward mechanism combining rule-based and model-based assessment, rather than specifically proposing a two-way substring exact match verifier for text-based QA tasks.

10. SR: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning

URL: View paper

Brief Assessment

Self-verify RL[62] focuses on rule-based verification methods for RL training but the provided context is too limited to determine if it specifically addresses two-way substring exact matching for general QA tasks. The available text only mentions 'rule-based' verification without sufficient detail to establish prior work on this specific verification approach.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] LoongRL: Reinforcement Learning for Advanced Reasoning over Long Contexts View paper
- [1] A survey of reinforcement learning for large reasoning models View paper
- [2] Spell: Self-play reinforcement learning for evolving long-context language models View paper
- [3] Llm post-training: A deep dive into reasoning large language models View paper

- [4] Learning Adaptive Parallel Reasoning with Language Models View paper
- [5] Kimi k1.5: Scaling Reinforcement Learning with LLMs View paper
- [6] Rlvmr: Reinforcement learning with verifiable meta-reasoning rewards for robust long-horizon agents View paper
- [7] Reinforcement learning foundations for deep research systems: A survey View paper
- [8] Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning View paper
- [9] Reinforcement Learning for Long-Horizon Interactive LLM Agents View paper
- [10] Effective Reinforcement Learning for Reasoning in Language Models View paper
- [11] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle View paper
- [12] Advancing Language Model Reasoning through Reinforcement Learning and Inference Scaling View paper
- [13] Theory of Mind for Multi-Agent Collaboration via Large Language Models View paper
- [14] WebExplorer: Explore and Evolve for Training Long-Horizon Web Agents View paper
- [15] Think-RM: Enabling Long-Horizon Reasoning in Generative Reward Models View paper
- [16] Reward Is Enough: LLMs Are In-Context Reinforcement Learners View paper
- [17] A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law View paper
- [18] Serving Long-Context LLMs at the Mobile Edge: Test-Time Reinforcement Learning-based Model Caching and Inference Offloading View paper
- [19] Context-lite multi-turn reinforcement learning for LLM agents View paper
- [20] Llms are in-context reinforcement learners View paper
- [21] JT-Math: A Multi-Stage Framework for Advanced Mathematical Reasoning in Large Language Models View paper
- [22] Memory-T1: Reinforcement Learning for Temporal Reasoning in Multi-session Agents View paper
- [23] MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent View paper
- [24] Learning to Reduce: Optimal Representations of Structured Data in Prompting Large Language Models View paper
- [25] QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning View paper
- [26] R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning View paper
- [27] SimpleVLA-RL: Scaling VLA Training via Reinforcement Learning View paper
- [28] Leveraging Large Language Model for Intelligent Log Processing and Autonomous Debugging in Cloud AI Platforms View paper
- [29] Reinforcement Fine-Tuning for Reasoning towards Multi-Step Multi-Source Search in Large Language Models View paper
- [30] Agentic Reinforced Policy Optimization View paper
- [31] E3-Rewrite: Learning to Rewrite SQL for Executability, Equivalence,and Efficiency View paper
- [32] Satori: Reinforcement Learning with Chain-of-Action-Thought Enhances LLM Reasoning via Autoregressive Search View paper
- [33] Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning View paper
- [34] The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models View paper
- [35] Beyond Numeric Rewards: In-Context Dueling Bandits with LLM Agents View paper
- [36] Contextual integrity in llms via reasoning and reinforcement learning View paper
- [37] Hierarchical DLO Routing with Reinforcement Learning and In-Context Vision-language Models View paper
- [38] Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning View paper
- [39] GFlowVLM: Enhancing Multi-step Reasoning in Vision-Language Models with Generative Flow Networks View paper
- [40] DeepPlanner: Scaling Planning Capability for Deep Research Agents via Advantage Shaping View paper
- [41] DeepTheorem: Advancing LLM Reasoning for Theorem Proving Through Natural Language and Reinforcement Learning View paper
- [42] Cache-Efficient Posterior Sampling for Reinforcement Learning with LLM-Derived Priors Across Discrete and Continuous Domains View paper
- [43] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning View paper
- [44] Scaling LLM Multi-turn RL with End-to-end Summarization-based Context Management View paper
- [45] LLM-Guided Reinforcement Learning for Interactive Environments View paper
- [46] Reinforcement learning: Advanced techniques for llm behavior optimization View paper
- [47] R-Search: Empowering LLM Reasoning with Search via Multi-Reward Reinforcement Learning View paper
- [48] Look back to reason forward: Revisitable memory for long-context llm agents View paper
- [49] QwenLong-L1.5: Post-Training Recipe for Long-Context Reasoning and Memory Management View paper
- [50] PilotRL: Training Language Model Agents via Global Planning-Guided Progressive Reinforcement Learning View paper
- [51] Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models View paper
- [52] Chain of agents: Large language models collaborating on long-context tasks View paper
- [53] The pokeagent challenge: Competitive and long-context learning at scale View paper
- [54] Retrieval-augmented hierarchical in-context reinforcement learning and hindsight modular reflections for task planning with llms View paper
- [55] Robohorizon: An llm-assisted multi-view world model for long-horizon robotic manipulation View paper
- [56] Amago: Scalable in-context reinforcement learning for adaptive agents View paper
- [57] Large language models are learnable planners for long-term recommendation View paper
- [58] Large Language Models Post-training: Surveying Techniques from Alignment to Reasoning View paper
- [59] VerIF: Verification Engineering for Reinforcement Learning in Instruction Following View paper
- [60] Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning View paper
- [61] Sparks of tabular reasoning via text2sql reinforcement learning View paper
- [62] SR: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning View paper
- [63] SATURN: SAT-based Reinforcement Learning to Unleash LLMs Reasoning View paper
- [64] DocThinker: Explainable Multimodal Large Language Models with Rule-based Reinforcement Learning for Document Understanding View paper
- [65] MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning View paper
- [66] WeThink: Toward General-purpose Vision-Language Reasoning via Reinforcement Learning View paper
- [67] DialogueReason: Rule-Based RL Sparks Dialogue Reasoning in LLMs View paper
- [68] Prompt-Based Length Controlled Generation with Reinforcement Learning View paper
- [69] Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks View paper

- [70] Plan-and-Act: Improving Planning of Agents for Long-Horizon Tasks View paper
- [71] What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices View paper
- [72] Odysseybench: Evaluating llm agents on long-horizon complex office application workflows View paper
- [73] Wildlong: Synthesizing realistic long-context instruction data at scale View paper
- [74] Generating Multi-turn Clarification for Web Information Seeking View paper
- [75] Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl View paper
- [76] Generalizing from short to long: Effective data synthesis for long-context instruction tuning View paper
- [77] Multi-Document Grounded Multi-Turn Synthetic Dialog Generation View paper