Novelty Assessment Report

Paper: Self-Aligned Reward: Towards Effective and Efficient Reasoners
PDF URL: https://openreview.net/pdf?id=89Pje8STvm
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Reinforcement learning with verifiable rewards has significantly advanced reasoning with large language models (LLMs) in domains such as mathematics and logic. However, verifiable signals provide only coarse-grained or binary correctness feedback. This limitation results in inefficiencies like overly verbose or repetitive reasoning. Existing length-based solutions (e.g., length penalty) compromise accuracy. To address this deficiency, we introduce **self-aligned reward (SAR)**, a generic, universally applicable self-guided signal that complements verifiable rewards to enhance both reasoning accuracy and efficiency in RL. Specifically, SAR is defined as the relative perplexity difference between an answer conditioned on the query and the standalone answer, thereby favoring responses that are concise and query-specific. Quantitative analysis reveals that SAR reliably judges answer quality: concise, correct answers score higher than redundant ones, and partially correct answers score higher than entirely incorrect ones. Evaluation on 4 different models across 7 benchmarks shows that integrating SAR with prevalent RL algorithms like PPO and GRPO reduces answer length by 30%, while improving accuracy by 4%. Our analysis also shows that SAR generalizes well to out-of-domain tasks and achieves a Pareto-optimal frontier between correctness and efficiency compared to state-of-the-art baselines. We also show that SAR shortens unnecessary elaboration while preserving advanced reasoning behaviors. These results highlight the promise of self-aligned reward as a fine-grained complement to verifiable rewards, paving the way for efficient and effective LLM training.

Core Task Landscape

This paper addresses: **Enhancing reasoning accuracy and efficiency in large language models through reinforcement learning**
A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

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

- **Reinforcement Learning Algorithms and Training Methods**
- **Reward Design and Verification Mechanisms**
- **Reasoning Capability Analysis and Evaluation**
- **Reasoning Architectures and Inference Strategies**
- **Domain-Specific Applications and Adaptations**
- **Survey and Review Literature**

Complete Taxonomy Tree

- Enhancing reasoning accuracy and efficiency in large language models through reinforcement learning Survey Taxonomy
- Reinforcement Learning Algorithms and Training Methods
  - Core RL Algorithm Development and Optimization (6 papers)
  - [5] Teaching large language models to reason with reinforcement learning (Havrilla, 2024) View paper
  - [22] Dapo: An open-source llm reinforcement learning system at scale (Yu, 2025) View paper
  - [26] Unlocking reasoning capabilities in llms via reinforcement learning exploration (Deng, 2025) View paper
  - [35] Large language models reasoning and reinforcement learning (Miquel Noguer I Alonso, 2023) View paper
  - [39] Effective Reinforcement Learning for Reasoning in Language Models (Li Shuo, 2025) View paper
  - [43] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't (Quy-Anh Dang, 2025) View paper
  - Training Dynamics and Stability Mechanisms (3 papers)
  - [8] The entropy mechanism of reinforcement learning for reasoning language models (Cui, 2025) View paper
  - [16] Emergent hierarchical reasoning in llms through reinforcement learning (Wang, 2025) View paper
  - [27] Demystifying long chain-of-thought reasoning in llms (Tong, 2025) View paper
  - Specialized Training Paradigms (5 papers)
  - [2] Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models (Liu Mingjie, 2025) View paper
  - [15] Reinforcement learning for reasoning in large language models with one training example (Wang Yi-ping, 2025) View paper
  - [18] Offline reinforcement learning for llm multi-step reasoning (Wang Huai-jie, 2025) View paper
  - [33] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models (Le, 2025) View paper
  - [34] Advancing language model reasoning through reinforcement learning and inference scaling (Hou Zhenyu, 2025) View paper
  - Non-Autoregressive and Alternative Generation Paradigms (2 papers)
  - [41] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning (Zhao Siyan, 2025) View paper
  - [50] Revolutionizing reinforcement learning framework for diffusion large language models (Wang Yinjie, 2025) View paper
- Reward Design and Verification Mechanisms
  - Self-Aligned and Efficiency-Oriented Rewards ★ (2 papers)
  - [0] Self-Aligned Reward: Towards Effective and Efficient Reasoners (Anon et al., 2026) View paper
  - [30] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning (Ning, 2025) View paper

- Verifiable Rewards and Rule-Based Verification (2 papers)
- [17] Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning (Xie Tian, 2025) View paper
- [49] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms (Liu Zihan, 2025) View paper
- Cross-Domain and Multi-Domain Reward Design (1 papers)
- [28] Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective (Cheng, 2025) View paper

- Reasoning Capability Analysis and Evaluation
  - Capability Boundary and Emergence Studies (3 papers)
  - [1] Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? (Yue Yang, 2025) View paper
  - [3] DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning (Daya Guo, 2025) View paper
  - [7] Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (DeepSeek-AI, 2025) View paper
  - Inference Scaling and Test-Time Reasoning (2 papers)
  - [10] Learning to reason with search for llms via reinforcement learning (M Chen, 2025) View paper
  - [14] Training language models to reason efficiently (Arora, 2025) View paper
- Reasoning Architectures and Inference Strategies
  - Search-Augmented and Interleaved Reasoning (3 papers)
  - [40] An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents (Jin Bo-wen, 2025) View paper
  - [42] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning (Li, 2025) View paper
  - [48] Interleaved Reasoning for Large Language Models via Reinforcement Learning (Qiu, 2025) View paper
  - Hierarchical and Multi-Model Reasoning Systems (1 papers)
  - [47] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router (Shao, 2025) View paper
- Domain-Specific Applications and Adaptations
  - Software Engineering and Code Evolution (1 papers)
  - [23] Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution (Wei, 2025) View paper
  - Formal Theorem Proving and Mathematical Reasoning (1 papers)
  - [21] Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning (Zhang Zi-yin, 2025) View paper
  - Information Retrieval and Document Ranking (1 papers)
  - [13] Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning (Zhuang, 2025) View paper
  - Multimodal Reasoning Applications (2 papers)
  - [6] Visionthink: Smart and efficient vision language model via reinforcement learning (Yang, 2025) View paper
  - [9] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models (Guanghao Zhou, 2025) View paper
  - Abstract Reasoning and Pattern Recognition (1 papers)
  - [38] AR2: Adversarial Reinforcement Learning for Abstract Reasoning in Large Language Models (Cheng-Kai Yeh, 2025) View paper
  - Non-NLP RL Applications (3 papers)
  - [12] Reinforcement Learning Problem Solving with Large Language Models (Gholamian, 2024) View paper
  - [20] Guiding pretraining in reinforcement learning with large language models (Du Yuqing, 2023) View paper
  - [45] DrugGen enhances drug discovery with large language models and reinforcement learning. (Mahsa Sheikholeslami, 2025) View paper
- Survey and Review Literature
  - RL for LLM Reasoning Surveys (5 papers)
  - [4] Towards large reasoning models: A survey of reinforced reasoning with large language models (Xu Fengli, 2025) View paper
  - [19] A technical survey of reinforcement learning techniques for large language models (Aggarwal, 2025) View paper
  - [24] Reasoning language models: A blueprint (Besta, 2025) View paper
  - [37] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle (Liu Ke-liang, 2025) View paper
  - [46] A survey of reinforcement learning for large reasoning models (Zhang Kai-Yan, 2025) View paper
  - General LLM Reasoning and Multi-Step Reasoning Surveys (4 papers)
  - [11] Advancing reasoning in large language models: Promising methods and approaches (Patil Avinash, 2025) View paper
  - [29] Reasoning with large language models, a survey (Aske Plaat, 2024) View paper
  - [31] Multi-step reasoning with large language models, a survey (Aske Plaat, 2025) View paper
  - [36] Llm post-training: A deep dive into reasoning large language models (Kumar, 2025) View paper
  - Domain-Specific Reasoning Surveys (1 papers)
  - [25] A survey on large language models for mathematical reasoning (Wang Peng-Yuan, 2025) View paper
  - LLM-Enhanced RL Surveys (1 papers)
  - [32] Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods (Yuji Cao, 2024) View paper
  - Knowledge Graph and Structured Reasoning Surveys (1 papers)
  - [44] A Collaborative Reasoning Framework Powered by Reinforcement Learning and Large Language Models for Complex Questions Answering over Knowledge Graph (Z Zhang, 2025) View paper

Narrative

Core task: Enhancing reasoning accuracy and efficiency in large language models through reinforcement learning. The field has organized itself around several complementary dimensions. One major branch focuses on Reinforcement Learning Algorithms and Training Methods, exploring how to adapt policy gradient techniques, offline RL, and multi-turn interactions (e.g., Multi-Turn RL[30]) to the unique challenges of language-based reasoning. A second branch examines Reward Design and Verification Mechanisms, investigating how to construct reliable feedback signals—whether through external verifiers, self-aligned objectives, or efficiency-oriented metrics—that guide models toward correct and concise reasoning traces. Meanwhile, Reasoning Capability Analysis and Evaluation studies probe what reasoning abilities emerge under RL training (RL Reasoning Capacity[1]), and Reasoning Architectures and Inference Strategies address structural choices such as search-based decoding, hierarchical planning, and interleaved reasoning-action loops. Domain-Specific Applications and Adaptations demonstrate how these methods transfer to mathematics, code generation, and multimodal settings, while Survey and Review Literature (Large Reasoning Models Survey[4], Reasoning LLMs Survey[29]) synthesizes progress across these branches.

Within Reward Design and Verification Mechanisms, a particularly active line of work contrasts external verification—where outcome correctness is checked by formal tools or human labels—with self-aligned and efficiency-oriented approaches that encourage models to internalize quality criteria and minimize redundant computation. Self-Aligned Reward[0] exemplifies this latter direction, proposing mechanisms that allow the model to refine its own reward signal without heavy reliance on external supervision, thereby reducing annotation costs and improving scalability. This emphasis on self-supervision and efficiency distinguishes it from works like DeepSeek-R1[3] or Teaching LLMs Reason[5], which often combine RL with more structured verification or distillation pipelines. The trade-off between external oversight and autonomous alignment remains a central open question, as researchers seek to balance the reliability of verifiable rewards with the flexibility and cost-effectiveness of self-aligned methods.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning

Authors: Ning, Yansong, Li Wei, Yansong NING, Fang Jun, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Compressing long chain-of-thought (CoT) from large language models (LLMs) is an emerging strategy to improve the reasoning efficiency of LLMs. Despite its promising benefits, existing studies equally compress all thoughts within a long CoT, hindering more concise and effective reasoning. To this end, we first investigate the importance of different thoughts by examining their effectiveness and efficiency in contributing to reasoning through automatic long CoT chunking and Monte Carlo rollouts. B...

Relationship Analysis

Both papers belong to the Self-Aligned and Efficiency-Oriented Rewards category, focusing on optimizing reasoning for both correctness and efficiency through reward design. While the original paper (Self-Aligned Reward) proposes a perplexity-based self-aligned reward that measures query-answer alignment to reduce verbosity while maintaining accuracy, the candidate paper (Not all thoughts are generated equal) takes a different approach by analyzing thought-level importance within long CoTs and proposing a collaborative framework where two LLMs (long-thought and short-thought) work together through multi-turn reinforcement learning. The key distinction is that the original uses a continuous self-judging signal based on perplexity differences, whereas the candidate employs thought chunking, Monte Carlo rollouts, and asynchronous policy optimization with separate specialized models.

Contributions Analysis

Overall novelty summary. The paper proposes Self-Aligned Reward (SAR), a self-guided signal based on relative perplexity differences that complements verifiable rewards to improve both reasoning accuracy and efficiency. It resides in the 'Self-Aligned and Efficiency-Oriented Rewards' leaf under 'Reward Design and Verification Mechanisms'. This leaf contains only two papers total, indicating a relatively sparse research direction within the broader taxonomy of fifty papers. The focus on balancing correctness with efficiency through self-supervision distinguishes this work from the more populated branches addressing core RL algorithms or verifiable correctness signals.

The taxonomy reveals neighboring leaves focused on 'Verifiable Rewards and Rule-Based Verification' and 'Cross-Domain and Multi-Domain Reward Design', both emphasizing external supervision or broader applicability rather than self-aligned efficiency. The parent branch 'Reward Design and Verification Mechanisms' contrasts external verification approaches—where correctness is checked by formal tools—with self-aligned methods that internalize quality criteria. The paper's emphasis on perplexity-based self-guidance positions it at the intersection of reward design and efficiency optimization, diverging from works that rely heavily on external verifiers or human feedback.

Among twenty-nine candidates examined, none clearly refute the three main contributions: SAR itself (ten candidates, zero refutable), integration with PPO/GRPO (nine candidates, zero refutable), and Pareto-optimal accuracy-efficiency trade-offs (ten candidates, zero refutable). This suggests that within the limited search scope, the specific mechanism of using relative perplexity as a self-aligned efficiency signal appears novel. However, the search scale is modest—top-K semantic matches plus citation expansion—and does not constitute an exhaustive review of all efficiency-oriented reward mechanisms in the broader RL-for-LLM literature.

Based on the limited literature search, the work appears to occupy a relatively unexplored niche combining self-supervision with efficiency optimization. The sparse population of its taxonomy leaf and absence of refuting candidates among twenty-nine examined papers suggest novelty, though the analysis does not cover all possible prior work on length penalties, perplexity-based rewards, or efficiency metrics in RL training. A more comprehensive search might reveal additional related efforts in adjacent research communities.

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

Contribution 1: Self-Aligned Reward (SAR)

Description: The authors propose a novel reward mechanism called Self-Aligned Reward that measures the relative perplexity difference between an answer conditioned on the query and the standalone answer. This self-guided signal provides fine-grained supervision beyond binary correctness, promoting concise and query-specific responses while maintaining reasoning accuracy.

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. BAMBINO-LM:(Bilingual-) Human-Inspired Continual Pretraining of BabyLM

URL: View paper

Brief Assessment

BAMBINO-LM[68] uses perplexity-based rewards in a PPO framework for bilingual language acquisition in small-scale models, not for improving reasoning accuracy and efficiency in large language models as SAR does. The technical contexts and objectives are fundamentally different.

2. Reasoner for Real-World Event Detection: Scaling Reinforcement Learning via Adaptive Perplexity-Aware Sampling Strategy

URL: View paper

Brief Assessment

Reasoner Event Detection[66] focuses on abnormal event detection in customer service dialogues using perplexity-aware sampling for curriculum learning, not on self-guided reward mechanisms for general reasoning tasks. The perplexity usage is fundamentally different: SAR uses relative perplexity difference between conditioned and standalone answers as a reward signal, while Reasoner Event Detection[66] uses perplexity to assess sample difficulty for adaptive sampling strategies.

3. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models

URL: View paper

Brief Assessment

CDE[62] focuses on curiosity-driven exploration using perplexity as an exploration bonus in RL, not as a self-aligned reward mechanism for measuring query-answer alignment. The perplexity usage differs fundamentally in purpose and formulation.

---

4. The good, the bad, and the hybrid: A reward structure showdown in reasoning models training

URL: View paper

Brief Assessment

Reward Structure Showdown[63] focuses on comparing hard (binary), continuous (multi-component), and hybrid reward structures for mathematical reasoning tasks. It does not propose or evaluate perplexity-based self-guided reward signals that measure relative perplexity differences between conditioned and standalone answers, which is the core novelty of SAR.

---

5. Gencls++: Pushing the boundaries of generative classification in llms through comprehensive sft and rl studies across diverse datasets

URL: View paper

Brief Assessment

GenCLS++[65] focuses on generative classification tasks using prompt engineering and RL with rule-based rewards for format and accuracy. It does not employ perplexity-based self-guided reward signals that measure relative perplexity differences between conditioned and standalone answers as proposed in the original paper's SAR mechanism.

---

6. Aligning Large Language Models from Self-Reference AI Feedback with one General Principle

URL: View paper

Brief Assessment

Self-Reference AI Feedback[70] focuses on AI-generated preference feedback for alignment using self-reference and general principles, not on perplexity-based reward mechanisms for reinforcement learning efficiency in reasoning tasks.

---

7. Decision-Making Large Language Model for Wireless Communication: A Comprehensive Survey on Key Techniques

URL: View paper

Brief Assessment

Decision-Making LLM Wireless[67] is a survey on LLMs for wireless communication applications, not a research contribution on reward mechanisms for reinforcement learning in language models. The retrieved fragments mention perplexity and reward signals only in passing, without proposing any self-guided reward mechanism based on relative perplexity differences.

---

8. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models

URL: View paper

Brief Assessment

Stepwise Perplexity Refinement[61] focuses on identifying and removing unimportant reasoning steps using perplexity changes to improve efficiency in chain-of-thought reasoning. This differs from SAR, which uses relative perplexity difference as a reward signal in reinforcement learning to promote concise, query-specific responses during training.

---

9. Delve into PPO: Implementation matters for stable RLHF

URL: View paper

Brief Assessment

Delve into PPO[69] focuses on implementation details and stability improvements for PPO in RLHF, not on novel reward mechanisms based on perplexity differences. The paper does not propose self-guided reward signals using relative perplexity measurements between conditioned and standalone answers.

---

10. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning

URL: View paper

Brief Assessment

Entropy-Performance Exchange[64] focuses on analyzing entropy dynamics across training stages and proposes reward shaping based on perplexity and positional information, but does not propose a self-aligned reward mechanism that measures relative perplexity difference between conditioned and standalone answers as the original paper does.

Contribution 2: Integration of SAR with RL algorithms (SA-PPO and SA-GRPO)

Description: The authors develop training methods that integrate Self-Aligned Reward with existing reinforcement learning algorithms PPO and GRPO, creating SA-PPO and SA-GRPO. These methods combine verifiable correctness signals with the self-aligned reward to achieve simultaneous improvements in both accuracy and efficiency.

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. Self-Rewarding PPO: Aligning Large Language Models with Demonstrations Only

URL: View paper

Brief Assessment

Self-Rewarding PPO[75] focuses on combining SFT and PPO using a coherent reward based on log policy ratios between SFT and pretrained models for demonstration-only alignment. The original paper integrates a self-aligned reward (based on perplexity differences) with verifiable correctness signals in PPO/GRPO for mathematical reasoning tasks. These are distinct reward mechanisms and application domains.

---

2. PPO-BR: Dual-Signal Entropy-Reward Adaptation for Trust Region Policy Optimization

URL: View paper

Brief Assessment

PPO-BR[74] focuses on adapting PPO's clipping threshold using entropy and reward signals for general RL environments (MuJoCo, Atari). The original paper integrates a self-aligned reward (perplexity-based) with PPO/GRPO specifically for language model reasoning tasks. These are fundamentally different reward mechanisms and application domains.

---

3. MedGround-R1: Advancing Medical Image Grounding via Spatial-Semantic Rewarded Group Relative Policy Optimization

URL: View paper

Brief Assessment

MedGround-R1[73] focuses on medical image grounding using spatial-semantic rewards for vision-language models, not on combining self-aligned reward with PPO/GRPO for language model reasoning tasks.

---

4. Reinforcement Learning for Large Language Model Fine-Tuning: A Systematic Literature Review

URL: View paper

Brief Assessment

RL Fine-Tuning Review[77] is a systematic literature review that surveys existing RL4LLM fine-tuning methods rather than proposing novel integration techniques. It does not present prior work that would refute the novelty of combining Self-Aligned Reward with PPO and GRPO algorithms.

---

5. Discriminative Policy Optimization for Token-Level Reward Models

URL: View paper

Brief Assessment

Discriminative Policy Optimization[72] focuses on token-level reward modeling through discriminative policies (Q-RM) for PPO/REINFORCE, not on combining multiple reward signals like verifiable correctness with self-aligned rewards in PPO/GRPO as described in the original contribution.

---

6. Automated Clinical Trial Data Analysis and Report Generation by Integrating Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) â¦

URL: View paper

Brief Assessment

Clinical Trial RAG[71] focuses on clinical trial data analysis using RAG and LLM, not on reinforcement learning algorithms or reward signal integration for language model training.

---

7. Optimizing Safe and Aligned Language Generation: A Multi-Objective GRPO Approach

URL: View paper

Brief Assessment

Multi-Objective GRPO[79] focuses on combining multiple reward signals (politeness, meaningfulness, actionability, safety) through a multi-label reward regression model for alignment tasks, not on integrating self-aligned reward (SAR) based on perplexity differences for reasoning efficiency and accuracy.

---

8. Goal-Directed Story Generation: Augmenting Generative Language Models with Reinforcement Learning

URL: View paper

Brief Assessment

Goal-Directed Story Generation[76] applies PPO to story generation tasks with goal-seeking objectives, not to mathematical/logical reasoning with multiple reward signals (verifiable + self-aligned). The domains and reward mechanisms are fundamentally different.

---

9. Causally-Enhanced Reinforcement Policy Optimization

URL: View paper

Brief Assessment

Causally-Enhanced Policy[78] focuses on causal coherence along the Z→X→Y pathway using Jacobian-based sensitivities and counterfactual hardening, not on combining multiple reward signals like self-aligned reward with verifiable correctness in PPO/GRPO.

Contribution 3: Demonstration of Pareto-optimal accuracy-efficiency trade-off

Description: The authors establish that their approach achieves a Pareto-optimal balance between reasoning accuracy and computational efficiency. Unlike existing length-based methods that sacrifice accuracy for efficiency, SAR simultaneously improves both metrics across multiple models and benchmarks.

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. syftr: Pareto-Optimal Generative AI

URL: View paper

Brief Assessment

syftr[56] focuses on Pareto-optimal tradeoffs in RAG pipeline configurations (balancing accuracy vs. cost), not on reasoning model training. The original paper addresses Pareto-optimality in LLM reasoning via reinforcement learning with self-aligned rewards.

---

2. An empirical analysis of compute-optimal inference for problem-solving with language models

URL: View paper

Brief Assessment

Compute-Optimal Inference[54] focuses on inference-time compute optimization through tree search methods (REBASE) for mathematical problem-solving, not on training-time RL optimization for reasoning. The Pareto-optimal trade-offs studied are fundamentally different contexts.

---

3. How Far Are We from Optimal Reasoning Efficiency?

URL: View paper

Brief Assessment

Optimal Reasoning Efficiency[53] focuses on establishing empirical reasoning efficiency frontiers through systematic benchmarking and introduces REG as a metric to quantify gaps from these frontiers. The original paper's SAR approach achieves Pareto-optimality through a self-aligned reward mechanism that simultaneously improves accuracy and efficiency, which is a different technical approach from the candidate's frontier-based optimization framework.

---

4. Aware First, Think Less: Dynamic Boundary Self-Awareness Drives Extreme Reasoning Efficiency in Large Language Models

URL: View paper

Brief Assessment

Dynamic Boundary Awareness[55] focuses on dynamic reasoning boundary self-awareness and adaptive length management based on model capabilities, not on establishing Pareto-optimal frontiers between accuracy and efficiency as a primary contribution.

---

5. Scaling laws for precision

URL: View paper

Brief Assessment

Scaling Laws Precision[58] focuses on precision-aware scaling laws for training and inference in language models, examining trade-offs between precision, parameters, and data. The original paper addresses Pareto-optimality between reasoning accuracy and computational efficiency through self-aligned reward mechanisms in reinforcement learning, which is a fundamentally different approach and domain.

---

6. Pareto multi-objective alignment for language models

URL: View paper

Brief Assessment

Pareto Multi-Objective[52] addresses multi-objective alignment in LLMs by balancing conflicting objectives (e.g., informativeness vs. conciseness) through Pareto optimization. The original paper focuses on achieving Pareto-optimal trade-offs specifically between reasoning accuracy and computational efficiency in reasoning tasks, which is a distinct application domain from the general multi-objective alignment problem.

---

7. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models

URL: View paper

Brief Assessment

Inference Scaling Laws[51] focuses on compute-optimal inference strategies (sampling, tree search) for mathematical problem-solving, not on training-time reward mechanisms for general reasoning tasks. The Pareto-optimality discussed in the candidate concerns inference-time compute budgets and model sizes, not training rewards that balance accuracy and efficiency.

---

8. Cost-efficient knowledge-based question answering with large language models

URL: View paper

Brief Assessment

Cost-Efficient QA[57] addresses a different problem domain (knowledge-based QA with model selection) rather than reasoning efficiency within a single model. The Pareto frontier in the candidate concerns selecting between different models (LLMs vs KGMs) to balance API costs and accuracy, not optimizing reasoning token length versus accuracy within one model's generation process.

---

9. Smartthinker: Learning to compress and preserve reasoning by step-level length control

URL: View paper

Brief Assessment

SmartThinker[60] focuses on step-level length control in reasoning chains rather than establishing Pareto-optimal frontiers between accuracy and efficiency. The candidate addresses a different technical problem (fine-grained step importance) compared to the original's system-level trade-off analysis.

---

10. Tabi: An efficient multi-level inference system for large language models

URL: View paper

Brief Assessment

Tabi[59] focuses on inference latency reduction for discriminative models using multi-level routing, not on training-time reinforcement learning trade-offs between reasoning accuracy and computational efficiency during model training.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 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. Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models

Detected in: Contribution: contribution_1

⚠ **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] Self-Aligned Reward: Towards Effective and Efficient Reasoners 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 incentivizes reasoning in LLMs through reinforcement learning View paper
- [4] Towards large reasoning models: A survey of reinforced reasoning with large language models View paper
- [5] Teaching large language models to reason with reinforcement learning View paper
- [6] Visionthink: Smart and efficient vision language model via reinforcement learning View paper
- [7] Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning View paper
- [8] The entropy mechanism of reinforcement learning for reasoning language models View paper
- [9] Reinforced mllm: A survey on rl-based reasoning in multimodal large language models View paper
- [10] Learning to reason with search for llms via reinforcement learning View paper
- [11] Advancing reasoning in large language models: Promising methods and approaches View paper
- [12] Reinforcement Learning Problem Solving with Large Language Models View paper

- [13] Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning View paper
- [14] Training language models to reason efficiently View paper
- [15] Reinforcement learning for reasoning in large language models with one training example View paper
- [16] Emergent hierarchical reasoning in llms through reinforcement learning View paper
- [17] Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning View paper
- [18] Offline reinforcement learning for llm multi-step reasoning View paper
- [19] A technical survey of reinforcement learning techniques for large language models View paper
- [20] Guiding pretraining in reinforcement learning with large language models View paper
- [21] Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning View paper
- [22] Dapo: An open-source llm reinforcement learning system at scale View paper
- [23] Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution View paper
- [24] Reasoning language models: A blueprint View paper
- [25] A survey on large language models for mathematical reasoning View paper
- [26] Unlocking reasoning capabilities in llms via reinforcement learning exploration View paper
- [27] Demystifying long chain-of-thought reasoning in llms View paper
- [28] Revisiting Reinforcement Learning for LLM Reasoning from A Cross-Domain Perspective View paper
- [29] Reasoning with large language models, a survey View paper
- [30] Not all thoughts are generated equal: Efficient llm reasoning via multi-turn reinforcement learning View paper
- [31] Multi-step reasoning with large language models, a survey View paper
- [32] Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods View paper
- [33] Reasoning Under 1 Billion: Memory-Augmented Reinforcement Learning for Large Language Models View paper
- [34] Advancing language model reasoning through reinforcement learning and inference scaling View paper
- [35] Large language models reasoning and reinforcement learning View paper
- [36] Llm post-training: A deep dive into reasoning large language models View paper
- [37] Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle View paper
- [38] AR2: Adversarial Reinforcement Learning for Abstract Reasoning in Large Language Models View paper
- [39] Effective Reinforcement Learning for Reasoning in Language Models View paper
- [40] An Empirical Study on Reinforcement Learning for Reasoning-Search Interleaved LLM Agents View paper
- [41] d1: Scaling Reasoning in Diffusion Large Language Models via Reinforcement Learning View paper
- [42] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning View paper
- [43] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't View paper
- [44] A Collaborative Reasoning Framework Powered by Reinforcement Learning and Large Language Models for Complex Questions Answering over Knowledge Graph View paper
- [45] DrugGen enhances drug discovery with large language models and reinforcement learning. View paper
- [46] A survey of reinforcement learning for large reasoning models View paper
- [47] Route-and-Reason: Scaling Large Language Model Reasoning with Reinforced Model Router View paper
- [48] Interleaved Reasoning for Large Language Models via Reinforcement Learning View paper
- [49] Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms View paper
- [50] Revolutionizing reinforcement learning framework for diffusion large language models View paper
- [51] Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models View paper
- [52] Pareto multi-objective alignment for language models View paper
- [53] How Far Are We from Optimal Reasoning Efficiency? View paper
- [54] An empirical analysis of compute-optimal inference for problem-solving with language models View paper
- [55] Aware First, Think Less: Dynamic Boundary Self-Awareness Drives Extreme Reasoning Efficiency in Large Language Models View paper
- [56] syftr: Pareto-Optimal Generative AI View paper
- [57] Cost-efficient knowledge-based question answering with large language models View paper
- [58] Scaling laws for precision View paper
- [59] Tabi: An efficient multi-level inference system for large language models View paper
- [60] Smartthinker: Learning to compress and preserve reasoning by step-level length control View paper
- [61] Stepwise perplexity-guided refinement for efficient chain-of-thought reasoning in large language models View paper
- [62] Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models View paper
- [63] The good, the bad, and the hybrid: A reward structure showdown in reasoning models training View paper
- [64] Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning View paper
- [65] Gencls++: Pushing the boundaries of generative classification in llms through comprehensive sft and rl studies across diverse datasets View paper
- [66] Reasoner for Real-World Event Detection: Scaling Reinforcement Learning via Adaptive Perplexity-Aware Sampling Strategy View paper
- [67] Decision-Making Large Language Model for Wireless Communication: A Comprehensive Survey on Key Techniques View paper
- [68] BAMBINO-LM:(Bilingual-) Human-Inspired Continual Pretraining of BabyLM View paper
- [69] Delve into PPO: Implementation matters for stable RLHF View paper
- [70] Aligning Large Language Models from Self-Reference AI Feedback with one General Principle View paper
- [71] Automated Clinical Trial Data Analysis and Report Generation by Integrating Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) â¦ View paper
- [72] Discriminative Policy Optimization for Token-Level Reward Models View paper
- [73] MedGround-R1: Advancing Medical Image Grounding via Spatial-Semantic Rewarded Group Relative Policy Optimization View paper
- [74] PPO-BR: Dual-Signal Entropy-Reward Adaptation for Trust Region Policy Optimization View paper
- [75] Self-Rewarding PPO: Aligning Large Language Models with Demonstrations Only View paper
- [76] Goal-Directed Story Generation: Augmenting Generative Language Models with Reinforcement Learning View paper
- [77] Reinforcement Learning for Large Language Model Fine-Tuning: A Systematic Literature Review View paper
- [78] Causally-Enhanced Reinforcement Policy Optimization View paper

- [79] Optimizing Safe and Aligned Language Generation: A Multi-Objective GRPO Approach View paper
- [80] Decoding-Time Language Model Alignment with Multiple Objectives View paper