Novelty Assessment Report

Paper: The Art of Scaling Reinforcement Learning Compute for LLMs
PDF URL: https://openreview.net/pdf?id=FMjeC9Msws
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Reinforcement learning (RL) has become central to training large language models (LLMs), yet the field lacks predictive scaling methodologies comparable to those established for pre-training. Despite rapidly rising compute budgets, there is no principled understanding of how to evaluate algorithmic improvements for scaling RL compute. We present the first large-scale systematic study, amounting to more than 400,000 GPU-hours, that defines a principled framework for analyzing and predicting RL scaling in LLMs. We fit sigmoidal compute-performance curves for RL training and ablate a wide range of common design choices to analyze their effects on asymptotic performance and compute efficiency. We observe: (1) Not all recipes yield similar asymptotic performance, Details such as loss aggregation, normalization, curriculum, and off-policy algorithm primarily modulate compute efficiency without materially shifting the asymptote, and (3) Stable, scalable recipes follow predictable scaling trajectories, enabling extrapolation from smaller-scale runs. Combining these insights, we propose a best-practice recipe, ScaleRL, and demonstrate its effectiveness by successfully scaling and predicting validation performance on a single RL run scaled up to 100,000 GPU-hours. Our work provides both a scientific framework for analyzing scaling in RL and a practical recipe that brings RL training closer to the predictability long achieved in pre-training.

Core Task Landscape

This paper addresses: **Predictive Scaling of Reinforcement Learning Compute for Large 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 Dynamics and Scaling Laws**
- **RL Algorithms and Training Methods for LLM Reasoning**
- **Infrastructure and Deployment Optimization**
- **Domain-Specific Applications and Integration**
- **Foundational Concepts and Survey Literature**

Complete Taxonomy Tree

- Predictive Scaling of Reinforcement Learning Compute for Large Language Models Survey Taxonomy
- RL Training Dynamics and Scaling Laws
  - Compute-Performance Scaling Laws for RL Post-Training ★ (3 papers)
  - [0] The Art of Scaling Reinforcement Learning Compute for LLMs (Anon et al., 2026) View paper
  - [11] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning (Tan Ze-lin, 2025) View paper
  - [12] Predictive Scaling Laws for Efficient GRPO Training of Large Reasoning Models (Bhargava, 2025) View paper
  - Pre-Training Scaling Laws and Compute Optimization (3 papers)
  - [5] Scaling laws and compute-optimal training beyond fixed training durations (Elie Bakouch, 2024) View paper
  - [25] The Journey Matters: Average Parameter Count over Pre-training Unifies Sparse and Dense Scaling Laws (Jin Tian, 2025) View paper
  - [26] LLMs on the Line: Data Determines Loss-to-Loss Scaling Laws (Mayilvahanan, 2025) View paper
  - Cross-Family and Multi-Benchmark Scaling Prediction (2 papers)
  - [15] Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo (Charles, 2025) View paper
  - [21] Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families (Polo, 2024) View paper
  - Quantization and Compression Scaling Effects (1 papers)
  - [35] Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens (Ge, 2024) View paper
- RL Algorithms and Training Methods for LLM Reasoning
  - Policy Gradient and Actor-Critic Methods (4 papers)
  - [1] Advancing language model reasoning through reinforcement learning and inference scaling (Zhenyu Hou, 2025) View paper
  - [7] Act Only When It Pays: Efficient Reinforcement Learning for LLM Reasoning via Selective Rollouts (Zheng Hai-zhong, 2025) View paper
  - [17] Offline Actor-Critic Reinforcement Learning Scales to Large Models (Springenberg, 2024) View paper
  - [34] Stabilizing Policy Gradients for Sample-Efficient Reinforcement Learning in LLM Reasoning (Melo, 2025) View paper
  - Offline and Iterative RL Methods (1 papers)
  - [40] RoiRL: Efficient, Self-Supervised Reasoning with Offline Iterative Reinforcement Learning (Sakhi, 2025) View paper
  - Reward Design and Verification Strategies (3 papers)

- [14] Exploring data scaling trends and effects in reinforcement learning from human feedback (Shen Wei, 2025) View paper
- [29] Incentivizing LLMs to Self-Verify Their Answers (Zhang Fuxiang, 2025) View paper
- [50] Reinforcing General Reasoning without Verifiers (Zhou Xiangxin, 2025) View paper
- Search and Inference Scaling Techniques (1 papers)
- [16] Towards System 2 Reasoning in LLMs: Learning How to Think With Meta Chain-of-Thought (Snell, 2025) View paper
- Multi-Task and Multi-Agent RL Frameworks (5 papers)
- [8] Heterogeneous Group-Based Reinforcement Learning for LLM-based Multi-Agent Systems (Chen Guan-zhong, 2025) View paper
- [20] Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling (Zhou Jia-ming, 2025) View paper
- [27] YOLO-MARL: You Only LLM Once for Multi-Agent Reinforcement Learning (Zhuang Yuan, 2024) View paper
- [32] Optima: Optimizing Effectiveness and Efficiency for LLM-Based Multi-Agent System (Weize Chen, 2024) View paper
- [46] MARLIN: Multi-Agent Reinforcement Learning with Murmuration Intelligence and LLM Guidance for Reservoir Management (Fu Heming, 2025) View paper
- Resource-Constrained and Small-Scale RL Training (2 papers)
- [6] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't (QA Dang, 2025) View paper
- [10] Scaling Up RL: Unlocking Diverse Reasoning in LLMs via Prolonged Training (Liu Mingjie, 2025) View paper

• Infrastructure and Deployment Optimization
- Dynamic Resource Allocation and Auto-Scaling (5 papers)
- [9] LLM-Cloud Complete: Leveraging Cloud Computing for Efficient Large Language Model-based Code Completion (Zhang Mingxuan, 2024) View paper
- [13] Dynamic Resource Allocation in Serverless ETL: AI-Driven Scaling and Cost Optimization Models (DN Rodrigues, 2025) View paper
- [28] Temporal Fusion Transformer Based Vertical Scaling Management for Kubernetes (Kemalcan Bora, 2025) View paper
- [30] Temporal-Aware GPU Resource Allocation for Distributed LLM Inference via Reinforcement Learning (Yu Zhiwei, 2025) View paper
- [31] Multi-Objective Reinforcement Learning for Resource-Optimal LLM Serving in SaaS Clouds (Naga, 2025) View paper
- Distributed and Serverless Inference Systems (3 papers)
- [3] Advancing serverless computing for scalable ai model inference: Challenges and opportunities (Li Wang, 2024) View paper
- [18] Adaptive layer splitting for wireless llm inference in edge computing: A model-based reinforcement learning approach (Chen Yuxuan, 2024) View paper
- [39] Deep Reinforcement Learning-Based Methods for Model Deployment in Edge Environments (Xifei Song, 2025) View paper
- Model Compression and Pruning for Deployment (1 papers)
- [33] FastForward Pruning: Efficient LLM Pruning via Single-Step Reinforcement Learning (Xin Yuan, 2025) View paper
- DevOps and CI/CD Pipeline Automation (3 papers)
- [22] A Review of Generative AI and DevOps Pipelines: CI/CD, Agentic Automation, MLOps Integration, and Large Language Models (Joshi, 2025) View paper
- [24] AI-Driven DevOps Automation for Cloud-Native Application Modernization (Mittal, 2025) View paper
- [41] A Generative AI Framework for Data Pipeline Optimization and Analytical Performance Enhancement (Narayanan, 2025) View paper

• Domain-Specific Applications and Integration
- Code Generation and Compiler Optimization (1 papers)
- [37] DeCOS: Data-Efficient Reinforcement Learning for Compiler Optimization Selection Ignited by LLM (Tianming Cui, 2025) View paper
- Autonomous Systems and Decision-Making (2 papers)
- [19] Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning (Xiancai Xiang, 2024) View paper
- [38] LEAD: LLM-enhanced deep reinforcement learning for stable decision-making in critical autonomous driving scenarios (Dongwei Xu, 2025) View paper
- Research Automation and Information Retrieval (1 papers)
- [4] DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments (Yuxiang Zheng, 2025) View paper
- Wireless Networks and Communication Systems (2 papers)
- [23] Deepseek-inspired exploration of rl-based llms and synergy with wireless networks: A survey (Qiao Yu, 2025) View paper
- [36] An overview of machine learning-enabled optimization for reconfigurable intelligent surfaces-aided 6g networks: From reinforcement learning to large language â¦ (H Zhou, 2024) View paper

• Foundational Concepts and Survey Literature (8 papers)
- [2] Reinforcement learning: Advanced techniques for llm behavior optimization (Hariharan, 2025) View paper
- [42] A Survey of Large Language Models - Foundations and Future Directions (Roffo, 2025) View paper
- [43] Understanding the Technical Foundations of Large Language Models: Architectures, Training, and Applications (Ediga, 2025) View paper
- [44] Learning from Within: Hidden-State Dynamics as Rewards for Training LLMs (Hajrullahu, 2025) View paper
- [45] Architecture Optimization and Data-Efficient Methods in Machine Learning (Wang, 2025) View paper
- [47] A survey of slow thinking-based reasoning LLMs using reinforcement learning and test-time scaling law (Qianjun Pan, 2025) View paper
- [48] Basics of Large Language Models - transformers to LLMs (Green, 2025) View paper
- [49] Emergent Abilities in Large Language Models: A Survey (Giorgi, 2025) View paper

Narrative

Core task: predictive scaling of reinforcement learning compute for large language models. The field structure reflects a multifaceted effort to understand and optimize how RL post-training scales with computational resources. The taxonomy organizes work into several main branches: RL Training Dynamics and Scaling Laws examines fundamental relationships between compute budgets and model performance, often through empirical studies of how reward signals and policy updates behave under varying resource allocations; RL Algorithms and Training Methods for LLM Reasoning focuses on algorithmic innovations such as policy gradient stabilization, selective rollout strategies, and novel reward formulations that improve sample efficiency; Infrastructure and Deployment Optimization addresses practical concerns like serverless inference architectures, GPU allocation strategies, and resource scheduling; Domain-Specific Applications and Integration explores how RL-enhanced LLMs are deployed in specialized contexts such as scientific research assistants

or real-time systems; and Foundational Concepts and Survey Literature provides broader context through reviews of LLM capabilities, emergent reasoning phenomena, and technical foundations. Representative works like Compute Optimal Training[5] and DiLoCo Scaling Laws[15] illustrate efforts to characterize training efficiency, while Predictive GRPO Laws[12] and Math Reasoning Scaling[11] probe how specific RL methods scale in reasoning-heavy domains.

Particularly active lines of work center on deriving predictive laws that relate compute investment to downstream task performance, balancing the trade-off between exploration costs and inference-time gains, and understanding when prolonged training yields diminishing returns. Scaling RL Compute[0] sits within the branch examining compute-performance scaling laws for RL post-training, closely aligned with Predictive GRPO Laws[12] and Math Reasoning Scaling[11], which similarly investigate how algorithmic choices and problem domains modulate scaling behavior. While Math Reasoning Scaling[11] emphasizes domain-specific benchmarks in mathematical problem-solving, Scaling RL Compute[0] takes a broader view of predictive modeling across diverse RL training regimes, aiming to forecast performance gains before committing large-scale resources. This contrasts with infrastructure-focused efforts like Serverless AI Inference[3] or deployment studies such as Edge Deployment RL[39], which prioritize operational efficiency over theoretical scaling predictions. The work contributes to an emerging consensus that principled resource allocation requires not only empirical scaling curves but also interpretable models that generalize across tasks and training configurations.

Related Works in Same Category

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

1. Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning

Authors: Tan Ze-lin, Yu, Xiaohang, Zhou Yi-Fan, He Qiang, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

While scaling laws for large language models (LLMs) during pre-training have been extensively studied, their behavior under reinforcement learning (RL) post-training remains largely unexplored. This paper presents a systematic empirical investigation of scaling behaviors in RL-based post-training, with a particular focus on mathematical reasoning. Based on 54 experiments across diverse model sizes and training settings, we characterize how model scale, data volume, and computational budget inter...

Relationship Analysis

Both papers belong to the same taxonomy category of deriving compute-performance scaling laws for RL post-training in LLMs. They overlap in studying how compute budget, model size, and training dynamics relate to RL performance through predictive scaling curves (sigmoidal in the original paper, power-law in the candidate). The key differences are: the original paper focuses on recipe optimization and design choices (loss aggregation, normalization, curriculum) to achieve predictable scaling with a unified ScaleRL recipe, while the candidate paper emphasizes empirical characterization of scaling behaviors across the Qwen2.5 model family (0.5B-72B), discovering learning efficiency saturation trends and data reuse strategies in constrained regimes.

---

2. Predictive Scaling Laws for Efficient GRPO Training of Large Reasoning Models

Authors: Bhargava, Vaishnavi, Ghosh, Rajat, George, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Fine-tuning large language models (LLMs) for reasoning tasks using reinforcement learning methods like Group Relative Policy Optimization (GRPO) is computationally expensive. To address this, we propose a predictive framework that models training dynamics and helps optimize resource usage. Through experiments on Llama and Qwen models (3B 8B), we derive an empirical scaling law based on model size, initial performance, and training progress. This law predicts reward trajectories and identifies th...

Relationship Analysis

Both papers belong to the same taxonomy category of deriving compute-performance scaling laws for RL post-training of LLMs, specifically focusing on predictive frameworks that relate compute budget to training performance. They overlap in their core goal of establishing predictive scaling relationships to optimize RL training efficiency and reduce computational waste. However, the original paper conducts a comprehensive 400,000+ GPU-hour empirical study across multiple design choices (loss types, normalization, curriculum, etc.) to develop ScaleRL as a general recipe with sigmoidal scaling curves, while the candidate paper focuses specifically on GRPO training dynamics with a narrower scope (3B-8B models, identifying three training phases) and emphasizes early stopping strategies rather than algorithmic recipe optimization.

Contributions Analysis

Overall novelty summary. The paper presents a predictive framework for RL scaling in LLMs using sigmoidal compute-performance curves, alongside a best-practice recipe called ScaleRL. It resides in the 'Compute-Performance Scaling Laws for RL Post-Training' leaf, which contains only three papers total, including this one. This represents a relatively sparse research direction within the broader taxonomy of 50 papers across 19 leaf nodes, suggesting the specific focus on predictive RL scaling laws for LLM post-training remains an emerging area with limited prior systematic investigation.

The taxonomy tree reveals that neighboring work concentrates on pre-training scaling laws (e.g., Compute Optimal Training, DiLoCo Scaling Laws) and cross-family prediction methods, but these explicitly exclude RL-specific post-training dynamics. The sibling papers in the same leaf—Predictive GRPO Laws and Math Reasoning Scaling—examine RL scaling in narrower contexts (specific algorithms or mathematical domains), whereas this work aims for broader predictive modeling across diverse RL training regimes. Adjacent branches address algorithmic innovations (policy gradient methods, reward design) and infrastructure optimization, but lack the systematic compute-performance prediction focus central to this contribution.

Among 13 candidates examined across three contributions, zero refutable pairs were identified. The predictive framework contribution examined one candidate with no refutation; the ScaleRL recipe examined two candidates with no refutation; and the comprehensive empirical study examined ten candidates with no refutation. This limited search scope—13 papers rather than an exhaustive review—suggests the analysis captures immediate semantic neighbors but may not reflect the full landscape of RL scaling research. The absence of refutations among examined candidates indicates that, within this bounded search, the specific combination of sigmoidal curve fitting, design choice ablations, and best-practice recipe formulation appears distinct from prior work.

Based on the limited literature search of 13 candidates, the work appears to occupy a relatively novel position within RL scaling law research, particularly in its systematic approach to predicting compute-performance trajectories. However, the sparse population of the taxonomy leaf and the constrained search scope mean this assessment reflects top-K semantic matches rather than comprehensive field coverage. The analysis does not capture potential overlaps with broader scaling law literature outside the examined candidate set.

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

Contribution 1: Predictive framework for RL scaling in LLMs using sigmoidal compute-performance curves

Description: The authors introduce a sigmoidal curve framework (Equation 1) that models the relationship between expected reward and training compute, enabling extrapolation of RL performance from lower-compute runs to higher compute budgets. This framework quantifies asymptotic performance (A) and compute efficiency (B), providing a predictive methodology for evaluating RL scalability.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Token-Efficient RL for LLM Reasoning

URL: View paper

Brief Assessment

Token Efficient RL[61] does not address predictive scaling frameworks or sigmoidal compute-performance curves. The candidate focuses on memory-efficient RL methods (S-GRPO and T-SPMO) for token-level optimization under LoRA constraints, not on modeling or extrapolating RL performance across compute budgets.

Contribution 2: ScaleRL: a best-practice RL recipe that scales predictably with compute

Description: The authors develop ScaleRL, an RL training recipe that integrates asynchronous Pipeline-RL, forced length interruptions, truncated importance sampling RL loss (CISPO), prompt-level loss averaging, batch-level advantage normalization, FP32 precision at logits, zero-variance filtering, and no-positive-resampling. This recipe achieves state-of-the-art asymptotic performance and compute efficiency while maintaining predictable scaling trajectories.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient Multi-turn RL for GUI Agents via Decoupled Training and Adaptive Data Curation

URL: View paper

Brief Assessment

GUI Agents[52] focuses on system-level efficiency for multi-turn GUI agent training with decoupled architecture, not on developing a general RL recipe with predictable scaling laws across compute budgets.

2. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm trainin

URL: View paper

Brief Assessment

LlamaRL[51] focuses on distributed system infrastructure and asynchronous training frameworks for large-scale LLM RL, not on developing algorithmic recipes that combine specific loss functions, normalization strategies, and curriculum methods like ScaleRL does.

Contribution 3: Comprehensive empirical study identifying three key principles for RL scaling

Description: Through over 400,000 GPU-hours of experiments, the authors systematically ablate design choices in RL training and establish three principles: different methods reach different performance ceilings, common interventions mainly affect compute efficiency rather than asymptotic performance, and scalable methods can be identified early by estimating scaling parameters from initial training dynamics.

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. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models

URL: View paper

Brief Assessment

Remax[60] focuses on algorithmic simplification of PPO for RLHF by exploiting properties like fast simulation and deterministic transitions, not on systematic scaling studies or compute-performance curves across different RL methods.

2. Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't

URL: View paper

Brief Assessment

Small LLM Reasoning[6] focuses on training a 1.5B parameter model under strict resource constraints (4 GPUs, 24 hours) with limited data (7,000 samples), examining training instability and length control issues. The ORIGINAL paper conducts a large-scale systematic study (400,000+ GPU-hours) establishing predictive scaling laws and three principles about asymptotic performance versus compute efficiency across different RL methods. These are fundamentally different scopes and objectives.

3. Teaching large language models to reason with reinforcement learning

URL: View paper

Brief Assessment

Teaching LLMs Reasoning[59] focuses on comparing RL algorithms (expert iteration, PPO, return-conditioned RL) for reasoning tasks with different reward schemes and initializations, but does not systematically study scaling laws, asymptotic performance ceilings, or compute efficiency parameters across different RL methods as the original paper does.

4. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning

URL: View paper

Brief Assessment

Test-time Compute Scaling[53] focuses on scaling inference-time computation for reasoning tasks, not on systematic ablation of RL training design choices or establishing principles for RL compute scaling during training. The candidate examines test-time compute allocation strategies (search vs. revisions), while the original paper studies RL training dynamics and asymptotic performance ceilings.

5. Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective

URL: View paper

Brief Assessment

Search and Learning Roadmap[58] provides a conceptual framework for reproducing o1 through RL components (policy initialization, reward design, search, learning) but does not present empirical ablation studies with fitted scaling curves or systematic identification of asymptotic performance versus compute efficiency principles as done in the original paper.

6. DistFlow: A Fully Distributed RL Framework for Scalable and Efficient LLM Post-Training

URL: View paper

Brief Assessment

DistFlow[57] focuses on distributed system architecture and infrastructure for RL training efficiency, not on empirical principles of RL scaling behavior. The candidate addresses engineering bottlenecks in multi-controller frameworks rather than studying asymptotic performance versus compute efficiency trade-offs in RL algorithms.

7. Is a Good Foundation Necessary for Efficient Reinforcement Learning? The Computational Role of the Base Model in Exploration

URL: View paper

Brief Assessment

Foundation for Exploration[55] focuses on computational-statistical tradeoffs in RL with language models through a theoretical sampling oracle framework, not on empirical scaling principles from large-scale experiments. The candidate addresses fundamentally different questions about coverage necessity and inference-time exploration efficiency.

8. Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning

URL: View paper

Brief Assessment

Math Reasoning Scaling[11] focuses on characterizing scaling laws through predictive power-law formulations relating test loss to compute/data, while the original paper establishes principles about asymptotic performance ceilings and compute efficiency through sigmoidal curve fitting. The candidate's emphasis is on mathematical reasoning scaling behaviors rather than general RL framework design principles.

9. Towards large reasoning models: A survey of reinforced reasoning with large language models

URL: View paper

Brief Assessment

Large Reasoning Models[54] is a survey paper that reviews existing work on reinforced reasoning with LLMs. It does not present original empirical studies on RL scaling principles or conduct ablation experiments on design choices.

10. Kimi k1. 5: Scaling reinforcement learning with llms

URL: View paper

Brief Assessment

Kimi k1.5[56] focuses on long-context scaling and policy optimization for RL with LLMs, but does not present a systematic empirical study with scaling curves or principles about asymptotic performance versus compute efficiency. The candidate emphasizes practical training recipes rather than predictive scaling frameworks.

Appendix: Text Similarity Detection

Textual similarity detection checked 14 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. Kimi k1. 5: Scaling reinforcement learning with llms

Detected in: Contribution: contribution_3

⚠ **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] The Art of Scaling Reinforcement Learning Compute for LLMs View paper
- [1] Advancing language model reasoning through reinforcement learning and inference scaling View paper
- [2] Reinforcement learning: Advanced techniques for llm behavior optimization View paper
- [3] Advancing serverless computing for scalable ai model inference: Challenges and opportunities View paper
- [4] DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments View paper
- [5] Scaling laws and compute-optimal training beyond fixed training durations View paper
- [6] Reinforcement Learning for Reasoning in Small LLMs: What Works and What Doesn't View paper
- [7] Act Only When It Pays: Efficient Reinforcement Learning for LLM Reasoning via Selective Rollouts View paper
- [8] Heterogeneous Group-Based Reinforcement Learning for LLM-based Multi-Agent Systems View paper
- [9] LLM-Cloud Complete: Leveraging Cloud Computing for Efficient Large Language Model-based Code Completion View paper
- [10] Scaling Up RL: Unlocking Diverse Reasoning in LLMs via Prolonged Training View paper
- [11] Scaling Behaviors of LLM Reinforcement Learning Post-Training: An Empirical Study in Mathematical Reasoning View paper
- [12] Predictive Scaling Laws for Efficient GRPO Training of Large Reasoning Models View paper
- [13] Dynamic Resource Allocation in Serverless ETL: AI-Driven Scaling and Cost Optimization Models View paper
- [14] Exploring data scaling trends and effects in reinforcement learning from human feedback View paper
- [15] Communication-Efficient Language Model Training Scales Reliably and Robustly: Scaling Laws for DiLoCo View paper
- [16] Towards System 2 Reasoning in LLMs: Learning How to Think With Meta Chain-of-Thought View paper
- [17] Offline Actor-Critic Reinforcement Learning Scales to Large Models View paper
- [18] Adaptive layer splitting for wireless llm inference in edge computing: A model-based reinforcement learning approach View paper
- [19] Real-time integration of fine-tuned large language model for improved decision-making in reinforcement learning View paper
- [20] Omni-Thinker: Scaling Multi-Task RL in LLMs with Hybrid Reward and Task Scheduling View paper
- [21] Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families View paper
- [22] A Review of Generative AI and DevOps Pipelines: CI/CD, Agentic Automation, MLOps Integration, and Large Language Models View paper
- [23] Deepseek-inspired exploration of rl-based llms and synergy with wireless networks: A survey View paper

- [24] AI-Driven DevOps Automation for Cloud-Native Application Modernization View paper
- [25] The Journey Matters: Average Parameter Count over Pre-training Unifies Sparse and Dense Scaling Laws View paper
- [26] LLMs on the Line: Data Determines Loss-to-Loss Scaling Laws View paper
- [27] YOLO-MARL: You Only LLM Once for Multi-Agent Reinforcement Learning View paper
- [28] Temporal Fusion Transformer Based Vertical Scaling Management for Kubernetes View paper
- [29] Incentivizing LLMs to Self-Verify Their Answers View paper
- [30] Temporal-Aware GPU Resource Allocation for Distributed LLM Inference via Reinforcement Learning View paper
- [31] Multi-Objective Reinforcement Learning for Resource-Optimal LLM Serving in SaaS Clouds View paper
- [32] Optima: Optimizing Effectiveness and Efficiency for LLM-Based Multi-Agent System View paper
- [33] FastForward Pruning: Efficient LLM Pruning via Single-Step Reinforcement Learning View paper
- [34] Stabilizing Policy Gradients for Sample-Efficient Reinforcement Learning in LLM Reasoning View paper
- [35] Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens View paper
- [36] An overview of machine learning-enabled optimization for reconfigurable intelligent surfaces-aided 6g networks: From reinforcement learning to large language â¦ View paper
- [37] DeCOS: Data-Efficient Reinforcement Learning for Compiler Optimization Selection Ignited by LLM View paper
- [38] LEAD: LLM-enhanced deep reinforcement learning for stable decision-making in critical autonomous driving scenarios View paper
- [39] Deep Reinforcement Learning-Based Methods for Model Deployment in Edge Environments View paper
- [40] RoiRL: Efficient, Self-Supervised Reasoning with Offline Iterative Reinforcement Learning View paper
- [41] A Generative AI Framework for Data Pipeline Optimization and Analytical Performance Enhancement View paper
- [42] A Survey of Large Language Models - Foundations and Future Directions View paper
- [43] Understanding the Technical Foundations of Large Language Models: Architectures, Training, and Applications View paper
- [44] Learning from Within: Hidden-State Dynamics as Rewards for Training LLMs View paper
- [45] Architecture Optimization and Data-Efficient Methods in Machine Learning View paper
- [46] MARLIN: Multi-Agent Reinforcement Learning with Murmuration Intelligence and LLM Guidance for Reservoir Management View paper
- [47] A survey of slow thinking-based reasoning LLMs using reinforcement learning and test-time scaling law View paper
- [48] Basics of Large Language Models - transformers to LLMs View paper
- [49] Emergent Abilities in Large Language Models: A Survey View paper
- [50] Reinforcing General Reasoning without Verifiers View paper
- [51] Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm trainin View paper
- [52] Efficient Multi-turn RL for GUI Agents via Decoupled Training and Adaptive Data Curation View paper
- [53] Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning View paper
- [54] Towards large reasoning models: A survey of reinforced reasoning with large language models View paper
- [55] Is a Good Foundation Necessary for Efficient Reinforcement Learning? The Computational Role of the Base Model in Exploration View paper
- [56] Kimi k1. 5: Scaling reinforcement learning with llms View paper
- [57] DistFlow: A Fully Distributed RL Framework for Scalable and Efficient LLM Post-Training View paper
- [58] Scaling of search and learning: A roadmap to reproduce o1 from reinforcement learning perspective View paper
- [59] Teaching large language models to reason with reinforcement learning View paper
- [60] Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models View paper
- [61] Token-Efficient RL for LLM Reasoning View paper