# DeepSeek-R1 Paper Review

## Enhancing LLM Reasoning through Reinforcement Learning

JEEON BAE
SSAFY 13th

February 25, 2025

# Outline

# Introduction

- **Background:** Recent studies have focused on enhancing LLM reasoning abilities using reinforcement learning.
- **Core Idea:**
  - **DeepSeek-R1-Zero:** Developed using pure RL without supervised fine-tuning.
  - **DeepSeek-R1:** Combines a small amount of cold-start data with multi-stage RL and SFT to improve readability and performance.
- **Evaluation:** Achieves competitive results on benchmarks such as AIME 2024, MATH-500, Codeforces, etc.

# Key Contributions

- **Pure RL-based Reasoning:** Demonstrates that LLM reasoning can be enhanced solely through reinforcement learning.
- **Utilization of Cold-Start Data:** Improves initial stability and readability by leveraging high-quality cold-start examples.
- **Multi-Stage Learning Pipeline:** Alternates between RL and SFT to maximize the overall performance of the model.
- **Distillation Technique:** Transfers effective reasoning patterns from large models to smaller ones.

# DeepSeek-R1-Zero: Pure RL Approach

- **RL Algorithm:** Utilizes Group Relative Policy Optimization (GRPO) to update the model without a critic.
- **Reward Modeling:** A description of the reward modeling process.
  - **Accuracy Reward:** Encourages the model to output answers in a specific format for rule-based evaluation.
  - **Format Reward:** Ensures that the chain-of-thought (CoT) is delimited, for example, by enclosing it within [THINK] and [/THINK] tags.
- **Self-Evolution:** The model naturally develops reflective and diverse problem-solving strategies during training.
- **Weakness:** Reduced readability & language mixing hindered practical use.

/

# GRPO

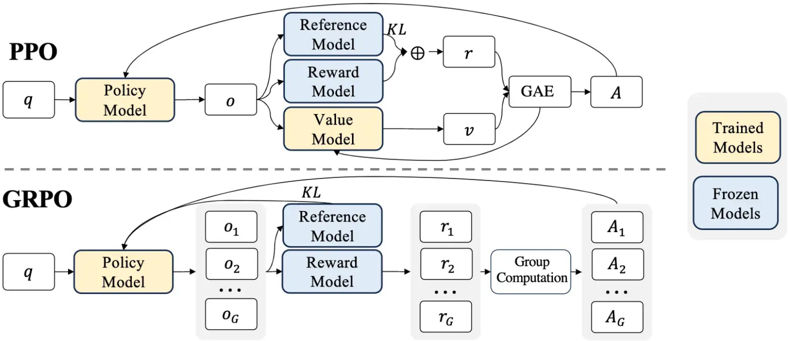
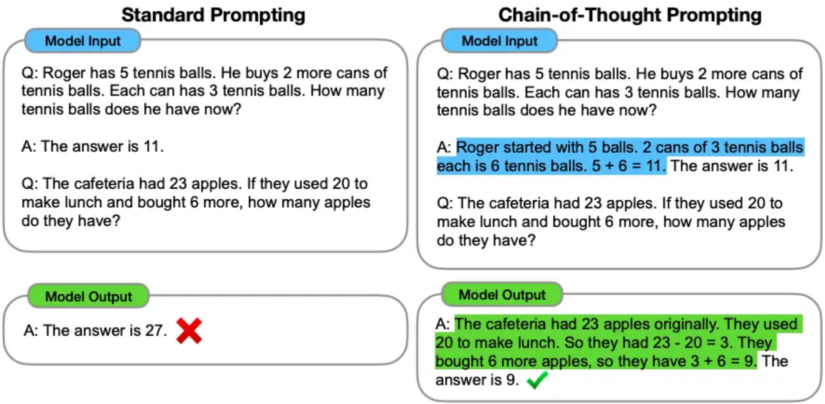



Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

# Chain-of-Thought



# Training Curve

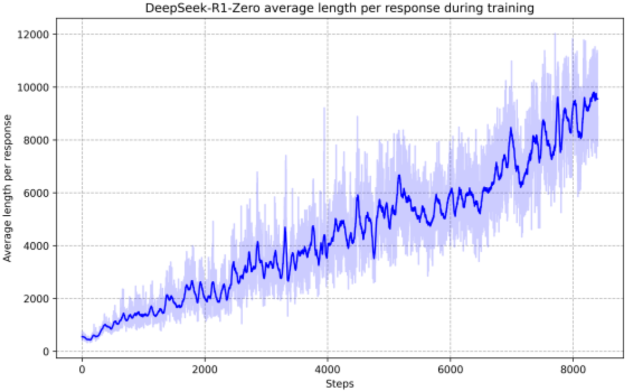
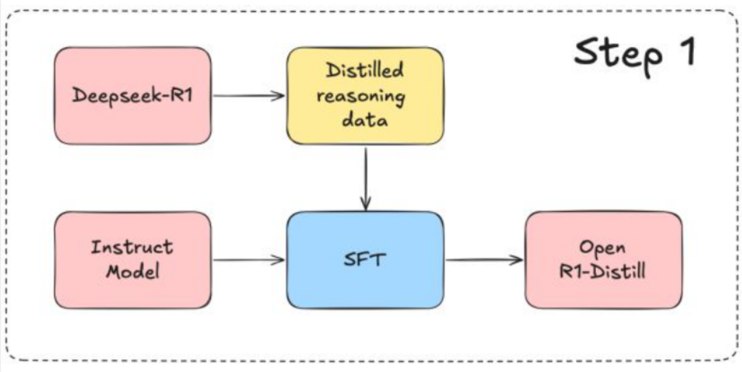



Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

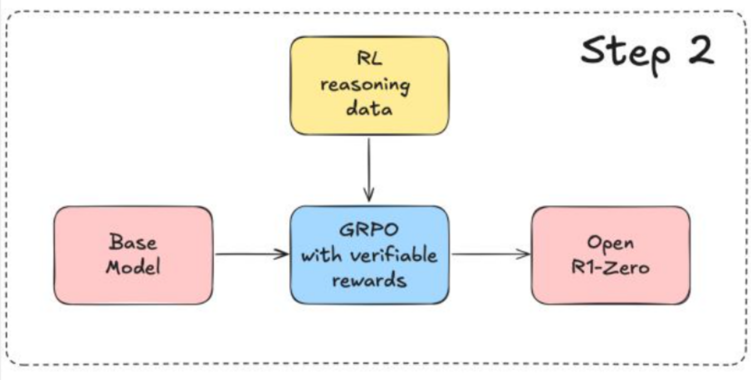
# DeepSeek-R1: Cold-Start and Multi-Stage Learning

- **Cold-Start Data:** Thousands of detailed CoT examples are used to fine-tune the base model before RL.
- **Integration of RL and SFT:**
  - An initial RL phase explores improved reasoning patterns.
  - Rejection sampling and subsequent SFT further refine the model.
- **Objective:** Achieve higher performance and better readability compared to pure RL methods.

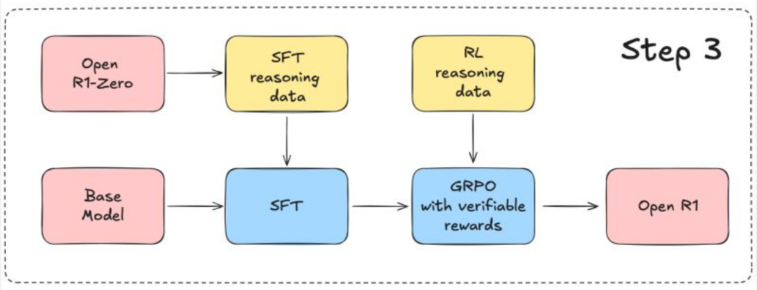
# Learning Pipeline



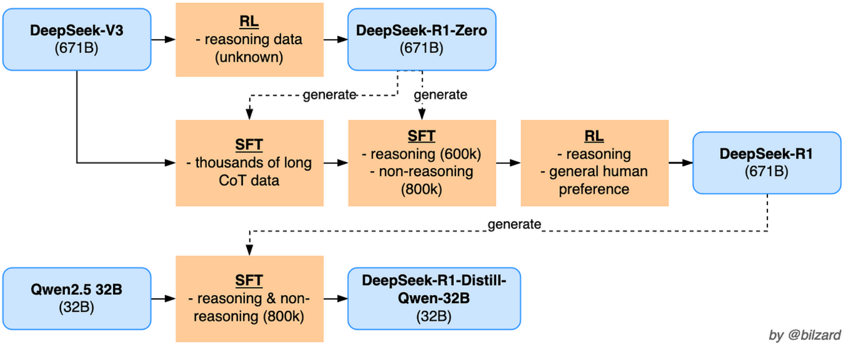
# Learning Pipeline



# Learning Pipeline



# Learning Pipeline

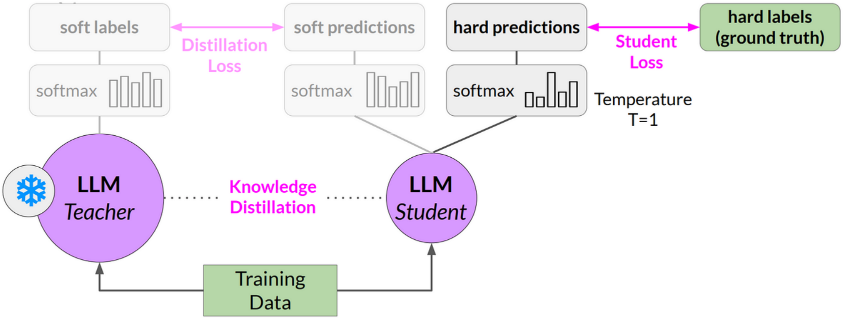


# Distillation: Transferring Reasoning to Smaller Models

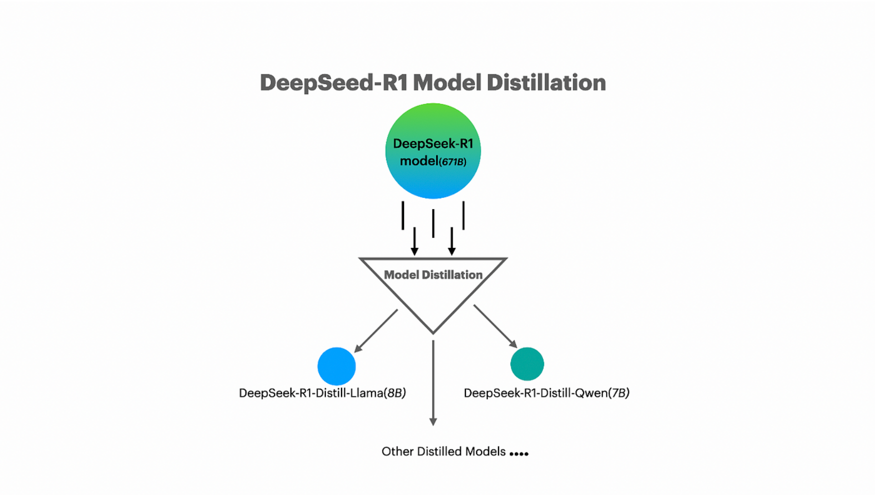
- **Goal:** Transfer effective reasoning patterns from DeepSeek-R1 to compact, dense models.
- **Target Models:** Models such as the Qwen2.5 and Llama series are distilled.
- **Results:** Distilled models (e.g., 14B, 32B, 70B) achieve superior performance on reasoning benchmarks.

# Distillation

Train a smaller student model from a larger teacher model



# Distillation



# Performance Comparison Chart of Distilled Models

| | AIME 2024 pass@1 | AIME 2024 cons@64 | MATH-500 pass@1 | GPQA Diamond pass@1 | LiveCodeBench pass@1 | CodeForces rating |
|---|---|---|---|---|---|---|
| GPT-4o-0513 | 9.3 | 13.4 | 74.6 | 49.9 | 32.9 | 759.0 |
| Claude-3.5-Sonnet-1022 | 16.0 | 26.7 | 78.3 | 65.0 | 38.9 | 717.0 |
| o1-mini | 63.6 | 80.0 | 90.0 | 60.0 | 53.8 | **1820.0** |
| QwQ-32B | 44.0 | 60.0 | 90.6 | 54.5 | 41.9 | 1316.0 |
| DeepSeek-R1-Distill-Qwen-1.5B | 28.9 | 52.7 | 83.9 | 33.8 | 16.9 | 954.0 |
| DeepSeek-R1-Distill-Qwen-7B | 55.5 | 83.3 | 92.8 | 49.1 | 37.6 | 1189.0 |
| DeepSeek-R1-Distill-Qwen-14B | 69.7 | 80.0 | 93.9 | 59.1 | 53.1 | 1481.0 |
| DeepSeek-R1-Distill-Qwen-32B | **72.6** | 83.3 | 94.3 | 62.1 | 57.2 | 1691.0 |
| DeepSeek-R1-Distill-Llama-8B | 50.4 | 80.0 | 89.1 | 49.0 | 39.6 | 1205.0 |
| DeepSeek-R1-Distill-Llama-70B | 70.0 | **86.7** | **94.5** | **65.2** | **57.5** | 1633.0 |

# Experimental Evaluation

- **Benchmarks:** Evaluated on AIME 2024, MATH-500, Codeforces, MMLU, etc.
- **Key Outcomes:**
  - AIME 2024: DeepSeek-R1 achieves a Pass@1 score of approximately 79.8%.
  - MATH-500: Pass@1 score of 97.3%.
  - Coding: Codeforces rating around 2029.
- **Comparisons:** Shows competitive performance against models such as OpenAI o1-1217.

# Performance

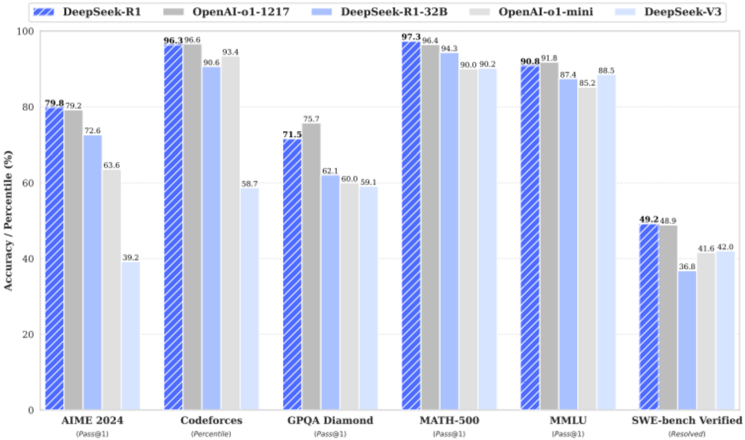



Figure 1 | Benchmark performance of DeepSeek-R1.

# Discussion

- **Strengths:**
  - Pure RL can effectively enhance LLM reasoning capabilities.
  - Distillation enables smaller models to achieve high reasoning performance.
- **Limitations:**
  - DeepSeek-R1-Zero suffers from readability and language-mixing issues.
  - High computational cost in the RL phase; limited improvements in some domains (e.g., software engineering).
- **Future Work:**
  - Broaden problem-solving abilities.
  - Address language-mixing issues and improve prompt engineering.
  - Enhance RL efficiency for software engineering tasks.

# Conclusion

- The DeepSeek-R1 series demonstrates a breakthrough in enhancing LLM reasoning via reinforcement learning.
- The combination of cold-start data with a multi-stage learning pipeline is key to improved performance and readability.
- Distillation techniques enable high reasoning performance even in compact models.
- Future research should explore broader applications and further efficiency improvements.

# Q & A

# References


- DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv:2501.12948v1.
- Shao et al., Group Relative Policy Optimization, 2024.

# The End