

# DeepSeek-R1 Reinforcement Learning Fine-Tuning and Mathematical Reasoning Benchmark Performance

Assignee Research

June 6, 2026

## Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of DeepSeek-R1's reinforcement learning fine-tuning on mathematical reasoning benchmarks relative to supervised fine-tuned baselines of similar parameter count. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: SED-SFT: Selectively Encouraging Diversity in Supervised Fine-Tuning. Research question: What is the impact of DeepSeek-R1's reinforcement learning fine-tuning on mathematical reasoning benchmarks relative to supervised fine-tuned baselines of similar parameter count?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

## 3 Results

13 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.5/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

Claim	Verified	Confidence
SED-SFT significantly enhances generation diversity with a negligible computational overhead increase compared with CE 1	✓	0.34
SED-SFT yields average improvements of 2.06 and 1.20 points in subsequent RL performance over standard CE-based baseline	✓	0.38
All experiments follow a standard SFT-then-RL pipeline to evaluate whether selectively encouraging diversity during SFT	×	0.13
We conduct experiments on two instruction-tuned backbones: Qwen2.5-Math-7B-Instruct and Llama-3.2-3B-Instruct.	×	0.12
For SFT, we sample 20,000 examples from the Micomind dataset.	×	0.03
For RL, we use the Math (Level 1) training split.	×	0.03
For SFT, we follow GEM’s training setup, using a learning rate of $2 \times 10^{-5}$ and DeepSpeed stage-2.	×	0.03
For RL, we apply GRPO implemented in the Verl framework with batch size 256; other hyperparameters use the default GRPO	×	0.04
We generate RL training samples by sampling each prompt 8 times with Qwen2.5-Math-7B-Instruct and filter out prompts whe	×	0.06
SED-SFT consistently outperforms the baselines.	×	0.11
The Cross-Entropy objective strongly drives the model policy $\pi_\theta$ to quickly converge along the single correct path $y_c$ , le	×	0.11
Updates should be strategically restricted to positions that potentially contain alternative reasoning branches to mitig	×	0.05

## References

- <http://arxiv.org/abs/2602.07464v1>
- <http://arxiv.org/abs/2509.25160v1>
- <http://arxiv.org/abs/2503.21776v4>