Few-Shot Prompting vs. Full Fine-Tuning in Noisy Reasoning Benchmarks for Large Language Models

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the accuracy degradation of large language models under few-shot prompting compare to full fine-tuning when evaluated on reasoning benchmarks with varying levels of label noise. Reinforcement learning (RL) has become a key technique for enhancing the reasoning abilities of large language models (LLMs), with policy-gradient algorithms dominating the post-training stage because of their efficiency and effectiveness. However, most existing benchmarks. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large Language Models Reasoning Abilities Under Non-Ideal Conditions After RL-Fine-Tuning. Research question: How does the accuracy degradation of large language models under few-shot prompting compare to full fine-tuning when evaluated on reasoning benchmarks with varying levels of label noise?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.3/10.

3 Results

10 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
The study investigates whether RL-fine-tuned LMs can perform summary inference when presented with diverse information.	×	0.07
The study investigates whether RL-fine-tuned LMs can ignore fine-grained noise to reach correct conclusions.	×	0.09
The study investigates whether RL-fine-tuned LMs can disregard irrelevant contextual information to reach valid conclusi	×	0.05
Qwen 2.5-VL-7B-Instruct was used as a Large Vision-Language Model baseline.	×	0.09
Llama 3.1-8B-Instruct, Qwen 3-14B, and Mistral-Small-24B-Instruct-2501 were used as Large Language Model baselines.	×	0.04
CommonsenseQA and Ceval-exam datasets were used to evaluate Research Question 1 for LLMs.	×	0.04
The CommonsenseQA dataset split used in the study contains 2000 training, 500 validation, and 1000 test samples.	×	0.02
The Ceval-exam dataset split used in the study contains 700 training, 246 validation, and 400 test samples.	×	0.02
Math12k, MathReasoning, Mathverse, and MathVision datasets were used to evaluate Research Questions 2 and 3.	×	0.03
For Research Question 2, TestA represents the original test set and FineTest represents the corresponding noisy test set	×	0.02
The study demonstrates that RL-fine-tuned LMs exhibit significant performance degradation under non-ideal scenarios comp	✓	0.19
The study proposes remediation strategies by manipulating format reward and example guidance.	×	0.01
The authors publicly released evaluation datasets designed to assess LM performance under noisy conditions.	×	0.03

References

- http://arxiv.org/abs/2308.10783v2
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2306.11066v2