Limitations of Language Model Benchmarks in Measuring Reasoning Capabilities

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v7. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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: What are the limitations of current language model evaluation benchmarks for measuring reasoning v7.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

12 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.7/10.

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	×	0.04
Qwen 2.5-VL-7B-Instruct was used as a Large Vision-Language Model baseline.	×	0.10
Llama 3.1-8B-Instruct, Qwen 3-14B, and Mistral-Small-24B-Instruct-2501 were used as Large Language Model baselines.	×	0.05
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.	×	0.01
The Ceval-exam dataset split used in the study contains 700 training, 246 validation, and 400 test samples.	×	0.01
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	×	0.03
For Research Question 3, TestB represents the original test set and FilterTest represents the corresponding test set wit	×	0.04
The study evaluated four RL-fine-tuned LMs and their variants across eight datasets.	×	0.06
RL-fine-tuned LMs exhibit significant performance degradation under non-ideal scenarios compared to ideal conditions.	✓	0.21
Remediation strategies were designed by manipulating format reward and example guidance.	×	0.04
The study publicly released evaluation datasets containing fine-grained distractors and irrelevant contextual informatio	×	0.04

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2407.04973v1