Limitations of Language Model Benchmarks in Measuring Reasoning Capabilities

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v14. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ARS: Adaptive Reasoning Suppression for Efficient Large Reasoning Language Models. Research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v14.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

14 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
ARS consistently achieves superior length reduction while maintaining competitive accuracy across all model scales.	×	0.11
ARS operates through three core components: (1) Multi-checkpoint certainty estimation, (2) Progressive threshold adaptat	✓	0.15
ARS establishes multiple checkpoints {c1, c2, . . . , ck} at regular intervals during generation.	×	0.02
At each checkpoint ci, ARS estimates model certainty through tentative answer probing.	×	0.05
The heuristic difficulty estimation is used to schedule the mode of operation in ARS.	×	0.02
ARS uses different policies (CoDFastPolicy, ElasticModeratePolicy, DeepReflectPolicy) based on the difficulty of the que	×	0.02
ARS sets a maximum token limit of 1200 tokens per response to prevent excessive generation.	×	0.03
ARS aims to minimize the expected output length E[T] while preserving reasoning accuracy.	×	0.06
ARS uses a loss function L to measure the difference between the extracted final answer f(o) and the ground truth y.	×	0.03
ARS uses specific keywords T = {"Wait", "But", "Alternatively", . . .} to identify reflection behaviors that often lead	×	0.05

References

- http://arxiv.org/abs/2505.19676v3
- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2407.04973v1