Limitations of Language Model Benchmarks in Measuring Reasoning Capabilities

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v8. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Wikipedia-based Semantic Interpretation for Natural Language Processing. Research question: What are the limitations of current language model evaluation benchmarks for measuring reasoning v8.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.0/10.

3 Results

4 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 7.0/10.

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
Adequate representation of natural language semantics requires access to vast amounts of common sense and domain-specifi	✓	0.38
Prior work in the field was based on purely statistical techniques that did not make use of background knowledge, on lim	✓	0.45
We propose a novel method, called Explicit Semantic Analysis (ESA), for fine-grained semantic interpretation of unrestri	✓	0.43
Our method represents meaning in a high-dimensional space of concepts derived from Wikipedia, the largest encyclopedia i	✓	0.35
We explicitly represent the meaning of any text in terms of Wikipedia-based concepts.	✓	0.32
Using ESA results in significant improvements over the previous state of the art in both text categorization and computi	✓	0.42
Due to the use of natural concepts, the ESA model is easy to explain to human users.	✓	0.31

References

- https://doi.org/10.1007/s10462-025-11253-3
- https://doi.org/10.1613/jair.2669
- https://doi.org/10.1145/3132747.3132748