Perplexity and Downstream Reasoning Performance in Language Models: A Meta-Analysis

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the relationship between language model perplexity and downstream reasoning task performance v6. 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: NeuralLog: Natural Language Inference with Joint Neural and Logical Reasoning. Research question: What is the relationship between language model perplexity and downstream reasoning task performance v6.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 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
The proposed NeuralLog model achieves state-of-the-art performance on the SICK dataset.	×	0.09
Removing the syntactic variation module (-ALBERT-SV) results in an 18.9 percentage point drop in accuracy.	×	0.05
Removing the monotonicity-based inference modules (-Monotonicity) results in a 15.6 percentage point drop in accuracy.	×	0.08
NeuralLog outperforms all deep learning-based baselines on the MED dataset by a significant amount.	×	0.11
NeuralLog performs better than BERT+ on upward inference by 15.4 percentage points.	×	0.04
NeuralLog performs better than BERT+ on downward inference by 23.6 percentage points.	×	0.04
NeuralLog shows a significant higher accuracy overall (+21.8) compared to BERT+ on the MED dataset.	×	0.04
The Natural Language Inference task can be modeled as a path planning problem.	✓	0.16
Monotonicity reasoning is a vertical action in the path planning model of NLI, where monotone inference moves up and sim	×	0.07
Syntactic variation and synonym replacement are horizontal actions in the path planning model of NLI, changing the form	×	0.03

References

- http://arxiv.org/abs/2402.04177v3
- http://arxiv.org/abs/2105.14167v3

- http://arxiv.org/abs/2601.01982v1