Scaling Laws for Language Model Performance in Logical Reasoning Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the effect of model size on language model performance on logical reasoning tasks v13. 9 claims were extracted from source literature; 0 were 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: LLMs as Models for Analogical Reasoning. Research question: What is the effect of model size on language model performance on logical reasoning tasks v13.

2 Methodology

Systematic literature search across multiple databases yielded 15 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

15 papers retrieved. 9 claims extracted; 0 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
Human subjects perform well overall in the Defaults setting, with a proportion of between 0.4 and 0.9 of answers matchin	×	0.03
The most advanced LLMs that we test match references in the range 0.1–0.95 in the different conditions we assess.	×	0.09
Results for LLMs (GPT-3, GPT-3.5, and GPT-4) across those conditions range from 0.1–1.0 in the two studies.	×	0.06
Increasing MMLU score is associated with more matches to reference answers on the Defaults condition.	×	0.02
Smaller models do not perform competitively (Pythia-12B obtains a reference match of 0.0, Falcon 40B 0.1, GPT-3 0.5, and	×	0.04
successful LLMs achieve human-level performance on our challenging tasks which require abstract rule induction and re-re	×	0.08
LLMs are more sensitive to presentation order and struggle when irrelevant semantic distractors are present.	×	0.06
The best performing LLM matched human performance across all conditions, and exceeded it in the compositional variants o	×	0.06
LLMs can achieve sophisticated analogical reasoning capabilities through domain-general learning mechanisms.	×	0.15

References

- http://arxiv.org/abs/2503.15113v1
- http://arxiv.org/abs/2406.13803v3
- http://arxiv.org/abs/2407.04973v1