Scaling Laws of Language Model Performance in Logical Reasoning Tasks

Assignee Research

June 4, 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. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ChaosBench-Logic: A Benchmark for Logical and Symbolic Reasoning on Chaotic Dynamical Systems. Research question: What is the effect of model size on language model performance on logical reasoning tasks.

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 3.8/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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 benchmark ground truth for each system is a truth assignment for the 11 predicates that is consistent with Φ.	×	0.04
The benchmark uses 11 unary predicates, each mapping a system s to a boolean.	×	0.03
The benchmark includes 30 dynamical systems, including continuous-time flows, discrete-time maps, PDEs, neuronal oscilla	×	0.06
The benchmark specifies a set Φ of global axioms encoding widely used implications in dynamical systems.	×	0.07
The benchmark uses the following annotation principles: Determinism vs. randomness, Chaos as a regime label, Predictabil	×	0.03
The benchmark introduces metrics that separate local correctness from global coherence.	×	0.07
The benchmark uses logical accuracy as a metric, defined as the fraction of correct predictions over Neval questions.	×	0.05
The benchmark avoids reverse implications to prevent overspecification and to force models to respect directionality.	×	0.02
The benchmark uses forward chaining under Φ to compute the correct answer for implication questions.	×	0.05
The benchmark uses a fixed ontology and ground truth to evaluate models on logical consistency.	×	0.05

References

- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2604.04177v2