Emergent Reasoning Capabilities in Transformers at Varying Model Scales

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the relationship between model scale and emergent reasoning capabilities in transformers v17. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Response: Emergent analogical reasoning in large language models. Research question: What is the relationship between model scale and emergent reasoning capabilities in transformers v17.

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 3.7/10.

3 Results

15 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
Humans achieved consistently higher accuracy than GPT-3 on modified letter string tasks involving a synthetic alphabet a	×	0.08
Human performance remains at a similar level across modifications involving synthetic alphabets and increased interval s	×	0.06
GPT-3 performance declines significantly for modified problem types involving synthetic alphabets and increased interval	×	0.05
GPT-3 generative accuracy for the synthetic alphabet is less than 0.1 on the tasks 'extend sequence', 'successor', 'pred	×	0.02
GPT-3 achieves accuracy on 'remove redundant letter' and 'sort' tasks similar to that reported in Webb, Holyoak, and Lu	×	0.12
GPT-3 accuracy is at least 30% of the original level on all counterfactual comprehension checks except for the 'predeces	×	0.05
Participants in the current study marginally outperformed participants in the original UCLA study on original tasks.	×	0.04
The human results represent the average performance of 121 UW undergraduate participants.	×	0.03
Each human participant received one randomly selected instance of each problem subtype.	×	0.03
GPT-3 results reflect the average performance across all 50 instances of the problems.	×	0.03
The synthetic alphabet was created by randomly changing the order of the letters in the real alphabet.	×	0.03
The synthetic alphabet was incorporated into tasks by preceding the original prompt with the sentence 'Use this fictiona	×	0.02
For problem types 'extend sequence', 'successor', and 'predecessor', the interval size for the letter to change was incr	×	0.03

References

- http://arxiv.org/abs/2308.16118v2
- http://arxiv.org/abs/2508.04848v1
- http://arxiv.org/abs/2407.04973v1