Emergent Reasoning in Transformers: Scaling Laws and Capability Thresholds

Assignee Research

June 7, 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 v18. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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 v18.

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 3.5/10.

3 Results

15 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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 performance remains at a level similar across modifications (Figure 4) while GPT-3 performance declines significan	×	0.08
The generative accuracy of GPT-3 for the synthetic alphabet is close to zero (< 0.1) when performing the modified tasks	×	0.04
Only for 'remove redundant letter' and 'sort' does GPT-3 achieve accuracy in a range similar to that reported in the ori	×	0.13
For all but on the 'precessor' task on the synthetic alphabet, we obtain a GPT-3 accuracy of at least 30% of the origina	×	0.04
Humans demonstrate significantly higher accuracy compared to GPT-3.	×	0.04
Human results represent the average performance of 121 participants (UW undergraduates).	×	0.03
Each participant received one randomly selected instance of each problem subtype.	×	0.04
GPT-3 results reflect the average performance across all 50 instances.	×	0.02
Gray error bars indicate 95% binomial confidence intervals for the average performance across multiple problems.	×	0.03
The subjects in our study marginally outperform those in the previous study, the similarity in performances is evidence	×	0.04
We create the synthetic alphabet by randomly changing the order of the letters in the real alphabet.	×	0.01
For both humans and GPT-3, we incorporate the synthetic alphabet in the tasks by preceding the original prompt with the	×	0.02
The increase in the size of the interval from one to two letters aims to rule out the possibility that GPT-3 merely repl	×	0.02

References

- http://arxiv.org/abs/2402.04177v3
- http://arxiv.org/abs/2308.16118v2
- http://arxiv.org/abs/2403.09832v1