Emergent Reasoning in Transformers as a Function of Model Scale

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the relationship between model scale and emergent reasoning capabilities in transformers. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Reasoning Effort and Problem Complexity: A Scaling Analysis in LLMs. Research question: What is the relationship between model scale and emergent reasoning capabilities in transformers.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

13 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
Success is assessed as a binary measure of whether the LLM outputs a valid solution to the Tents puzzle instance adherin	×	0.05
Problem complexity is quantified by problem size, defined as the product of grid dimensions (rows $\times$ columns).	×	0.10
Reasoning effort is measured by the total number of tokens generated by the LLMs, including all thinking tokens.	×	0.06
The study evaluated Gemini 2.0 Flash Thinking, OpenAI o3-mini, DeepSeek R1, and Qwen/QwQ-32B-Preview.	×	0.03
The linear fit for DeepSeek R1 reasoning effort versus problem size has an R2 score of 0.667.	×	0.08
The linear fit for o3-mini reasoning effort versus problem size has an R2 score of 0.833.	×	0.08
The linear fit for Qwen/QwQ-32B-Preview reasoning effort versus problem size has an R2 score of 0.087.	×	0.07
Gemini 2.0 Flash Thinking was excluded from the reasoning effort scaling analysis due to an unknown number of thinking t	×	0.08
No model solved problems larger than size 100.	×	0.03
o3-mini achieves the highest success rate among the evaluated models.	×	0.03
Qwen/QwQ-32B-Preview struggles with problem instances larger than size 20.	×	0.02
DeepSeek R1 consistently uses more tokens than o3-mini for successfully solved puzzles.	×	0.03
The linear fit for 'medium' difficulty reasoning effort versus problem size has an R2 score of 0.833.	×	0.08
The linear fit for 'high' difficulty reasoning effort versus problem size has an R2 score of 0.813.	×	0.08
The linear fit for 'low' difficulty reasoning effort versus problem size has an R2 score of 0.489.	×	0.08
The linear fit for 'easy' difficulty reasoning effort versus problem size has an R2 score of 0.468.	×	0.08
The linear fit for 'tricky' difficulty reasoning effort versus problem size has an R2 score of 0.502.	×	0.08

References

- http://arxiv.org/abs/2503.15113v1
- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2308.16118v2