Emergent Reasoning in Transformers at Scale: A Multi-Study Synthesis

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the relationship between model scale and emergent reasoning capabilities in transformers v12. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: How Does Critical Batch Size Scale in Pre-training?. Research question: What is the relationship between model scale and emergent reasoning capabilities in transformers v12.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

4 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
Setting the warmup steps according to the heuristic and using the ratio proportionally for all other model sizes is effe	×	0.02
Applying Exponentially Weighted Averages (EWA) can help smooth out noise, allowing the optimization to converge to the t	×	0.00
A very high EWA decay rate would be needed even for a 1.2B model with a moderate batch size of 1024.	×	0.07
By step 10,000, most runs achieve a validation loss below 3.2 (Figure 8a), and similarly, a loss below 2.8 is reached by	×	0.01
To reach the target loss of 2.736, the difference between the best and second-best runs grows substantially, with the be	×	0.00
The Constant+EWA scheduler performs competitively with cosine scheduling and outperforms WSD scheduling, especially for	×	0.07
Under small batch sizes, the schedule-free optimizer is a competitive baseline but it is significantly worse for batch s	×	0.03
Longer training requires higher EWA decay rate τ.	×	0.05
Training with longer duration may require a lower learning rate as suggested in DeepSeek-AI et al., 2024.	×	0.05
In the Chinchilla setting, keeping the data-to-model size ratio D/N = CChin constant, CBS increases with scale.	×	0.10
When controlling for either model size or data size, the growth in target losses becomes mostly dependent on data size r	×	0.12

References

- http://arxiv.org/abs/2410.21676v4

- http://arxiv.org/abs/2308.16118v2
- http://arxiv.org/abs/2504.16021v1