Density-Based Data Augmentation and Scaling Laws in Low-Resource Language Model Reasoning

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Does integrating density-based data augmentation into pre-training pipelines improve the scaling laws of language models on low-resource reasoning tasks compared to standard mixing strategies. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Analysing The Impact of Sequence Composition on Language Model Pre-Training. Research question: Does integrating density-based data augmentation into pre-training pipelines improve the scaling laws of language models on low-resource reasoning tasks compared to standard mixing strategies?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.2/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.2/10.

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.

References

- http://arxiv.org/abs/2411.15497v3
- http://arxiv.org/abs/2402.13991v1
- http://arxiv.org/abs/2412.10008v1