Pretraining Data Quality and Its Impact on Language Model Reasoning Performance

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does pretraining data quality affect language model reasoning benchmark performance v7. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Procedural Pretraining: Warming Up Language Models with Abstract Data. Research question: How does pretraining data quality affect language model reasoning benchmark performance v7.

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 4.2/10.

3 Results

12 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
Procedural pretraining improves performance and accelerates language model pretraining.	×	0.08
Procedural pretraining suggests the promise of disentangling knowledge acquisition from reasoning in LLMs.	×	0.07
Code is available at this page.	×	0.06
Procedural pretraining speeds up standard pretraining and improves performance on diverse domains.	×	0.08
Different pretrained layers (MLP vs. attention) contribute differently to each domain.	×	0.05
The pretrained information is localized in specific layers (attention vs. MLPs).	×	0.03
Procedural pretraining improves over standard pretraining with as little as 0.1 – 0.3% extra procedural tokens.	×	0.09
Procedural data is an efficient substitute to standard data, enabling models to reach the same loss with 55%, 67%, and 8	✓	0.24
Procedural pretraining gains persist on downstream language, code generation, and commonsense reasoning tasks.	×	0.09
Procedural pretraining facilitates learning different algorithms.	×	0.09
Shuffling the sequences of procedural data reduces performance.	×	0.09

References

- http://arxiv.org/abs/1312.3005v3
- http://arxiv.org/abs/2504.19565v3
- http://arxiv.org/abs/2601.21725v2