Pretraining Data Quality and Its Impact on Language Model Reasoning Performance

Assignee Research

June 6, 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 v17. 15 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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 v17.

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 4.8/10.

3 Results

12 papers retrieved. 15 claims extracted; 3 independently verified. Quality review score: 4.8/10.

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 compared to standard pretraining	×	0.09
Procedural pretraining consistently improves over standard pretraining using only 0.1% to 0.3% extra procedural tokens.	×	0.07
On the C4 dataset, procedural pretraining enables models to reach the same loss as standard pretraining using 55% of the	✓	0.17
On the CODEPARROT dataset, procedural pretraining enables models to reach the same loss as standard pretraining using 67	✓	0.17
On the DEEPMIND-MATH dataset, procedural pretraining enables models to reach the same loss as standard pretraining using	✓	0.20
The study validates findings across model sizes up to 1.3 billion parameters.	×	0.03
The study validates findings across data sizes up to 10.5 billion tokens.	×	0.04
Gains from procedural pretraining persist on downstream language, code generation, and commonsense reasoning tasks.	×	0.10
Pretrained information from procedural data is localized in specific layers, with MLPs and attention mechanisms contribu	×	0.08
Different types of procedural pretraining facilitate learning different algorithmic skills.	×	0.10
Shuffling the sequences of procedural data reduces performance compared to using structured procedural data.	×	0.09
The best procedural data type for the HAYSTACK task is 16 DYCK.	×	0.07
The best procedural data type for the REVERSED ADDITION task is ECA.	×	0.04
The best procedural data type for the MULTIPLICATION task is UNION and DELETE.	×	0.04
The best procedural data type for the SORTING task is 16 DYCK and SET.	×	0.05

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2601.21725v2
- http://arxiv.org/abs/2407.04973v1