Procedural Pretraining Data Quality and Language Model Reasoning Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does pretraining data quality affect language model reasoning benchmark performance v14. 10 claims were extracted from source literature; 2 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 v14.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 10 claims extracted; 2 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.	×	0.08
Procedural pretraining can be complementary to standard pretraining datasets, improving performance with as little as 0.	×	0.08
Procedural data enables models to reach the same loss with 55% of the original data on C4, 67% on CODEPARROT, and 86% on	✓	0.22
Procedural pretraining gains persist on downstream language, code generation, and commonsense reasoning tasks.	×	0.08
Different types of procedural pretraining facilitate learning different algorithmic skills.	×	0.10
The pretrained information is localized in specific layers (attention vs. MLPs).	×	0.04
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, reducing the required data size significantly.	×	0.09
Procedural pretraining improves performance on diverse domains, including natural language, code, and informal mathemati	✓	0.16
Procedural pretraining is validated across different model sizes (up to 1.3B parameters) and data sizes (up to 10.5B tok	×	0.07

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2601.21725v2
- http://arxiv.org/abs/2504.19565v3