Procedural Pretraining Effects on Model Alignment and Benchmark Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does pretraining on procedural data influence alignment metrics like toxicity and helpfulness in models evaluated on benchmarks like TruthfulQA and HELM. 7 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. 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 on procedural data influence alignment metrics like toxicity and helpfulness in models evaluated on benchmarks like TruthfulQA and HELM?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

10 papers retrieved. 7 claims extracted; 1 independently verified. Quality review score: 5.0/10.

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
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.	×	0.09
Procedural data enables models to reach the same loss with 55% of the original data on C4, 67% on CODEPARROT, and 86% on	✓	0.23
Procedural pretraining shows gains across different model sizes (up to 1.3B parameters) and data sizes (up to 10.5B toke	×	0.08
Procedural pretraining improves downstream language, code generation, and commonsense reasoning tasks.	×	0.09
Different types of procedural pretraining facilitate learning different algorithmic skills.	×	0.10
Shuffling the sequences of procedural data reduces performance, indicating the importance of sequence order in procedura	×	0.08

References

- http://arxiv.org/abs/2310.00905v2
- http://arxiv.org/abs/2601.21725v2
- http://arxiv.org/abs/2309.02144v1