Pretraining Data Quality and Its Impact on Language Model Reasoning Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does pretraining data quality affect language model reasoning benchmark performance v11. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation. Research question: How does pretraining data quality affect language model reasoning benchmark performance v11.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

13 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
The MR-Score metric consists of three sub-metrics: Matthews Correlation Coefficient (MCC), accuracy of the first-error-s	×	0.03
The MCC score ranges from -1 to +1, where -1 indicates total disagreement between prediction and observation, 0 suggests	×	0.03
The evaluated models were tested under both zero-shot and few-shot settings.	×	0.03
The inference temperature was set to zero across all models to ensure reproducibility and minimize variance.	×	0.02
The models evaluated include Qwen-v1.5-1.8B, Llama3-70B, Deepseek-v2-236B, WizardMath-v1.1-7B, MAmmoTH-70B, DeepseekMath	×	0.06
In the context of this paper, negative values are interpreted as no better than random guesses, and 0 is set as the cut-	×	0.03
The MR-Score for Qwen-1.8B under zero-shot setting is 0.1.	×	0.02
The MR-Score for Phi3-3.8B under few-shot setting is 21.9.	×	0.02
The MR-Score for Llama3-70B under zero-shot setting is 38.3.	×	0.02
The MR-Score for Deepseek-v2-236B under few-shot setting is 34.1.	×	0.06
The MR-Score for GPT-4-Turbo under zero-shot setting is 50.5.	×	0.03

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2409.04556v2
- http://arxiv.org/abs/2312.17080v4