Pre-Training Data Composition and Cross-Lingual Code Generation Performance in MultiPL-E

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the correlation between pre-training data composition and cross-language code generation performance across the 18 languages in the MultiPL-E benchmark. 6 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Cross-Lingual Pitfalls: Automatic Probing Cross-Lingual Weakness of Multilingual Large Language Models. Research question: What is the correlation between pre-training data composition and cross-language code generation performance across the 18 languages in the MultiPL-E benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.7/10.

3 Results

16 papers retrieved. 6 claims extracted; 3 independently verified. Quality review score: 6.7/10.

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
Linguistically related languages share similar performance patterns and benefit from targeted post-training.	✓	0.27
English is the primary training language for LLMs, and they generally perform best in English.	×	0.05
Cross-lingual weakness is defined as a model answering correctly in English but incorrectly in at least one other langua	×	0.11
The proposed beam search-based methodology efficiently uncovers cross-lingual weaknesses in LLMs.	✓	0.24
The performance scores for different languages and models are provided in the benchmark tables.	×	0.04
The correlation coefficients between linguistic similarity and cross-lingual weaknesses are provided for Chinese, Japane	✓	0.16

References

- http://arxiv.org/abs/2501.02338v1
- http://arxiv.org/abs/2505.18673v1
- http://arxiv.org/abs/2105.12485v2