CAKE and CodeLlama Performance in MultiPL-E Cross-Language Code Generation

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the CAKE model's performance in cross-language code generation tasks (e.g., Python to Java) compare to specialized code LLMs like CodeLlama when evaluated on MultiPL-E's test-driven. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MultiPL-E: A Scalable and Extensible Approach to Benchmarking Neural Code Generation. Research question: How does the CAKE model's performance in cross-language code generation tasks (e.g., Python to Java) compare to specialized code LLMs like CodeLlama when evaluated on MultiPL-E's test-driven benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
HumanEval is a diverse collection of 164 problems, where all problems have tests to check correctness, and most have exa	×	0.02
The best model evaluated by Fried et al. achieves only a 36% pass rate on Python.	×	0.02
MBPP is another large, commonly used benchmark of Python problems.	×	0.07
MultiPL-E supports 19 programming languages, which we categorize into four frequency classes (NICHE, LOW, MEDIUM, or HIG	×	0.06
Eight of the languages in MultiPL-E had never been used before to measure NL2Code performance.	×	0.05
MultiPL-E translates unit tests, doctests, Python-specific terminology, and type annotations.	×	0.04
MultiPL-E provides two parallel benchmarks for code generation in 19 languages encompassing a variety of programming par	✓	0.19
MultiPL-E includes a multi-language parallel evaluation of three models, Codex, InCoder, and CodeGen.	✓	0.15
MultiPL-E explores language frequency effects, the impact of type annotations, and prompt translation sensitivity on cod	×	0.11
The MultiPL-E system, dataset, and tutorial are available at github.com/nuprl/MultiPL-E.	×	0.03

References

- `http://arxiv.org/abs/2208.08227v4`

- http://arxiv.org/abs/2509.22472v1
- http://arxiv.org/abs/2401.10065v3