Cross-Lingual CodeT5 Consistency Regularization and Execution Accuracy on MBPP

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of model consistency regularization on the execution accuracy of cross-lingual CodeT5 models on the MBPP Python benchmark when trained with Gaussian noise augmentation across. 8 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. 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 impact of model consistency regularization on the execution accuracy of cross-lingual CodeT5 models on the MBPP Python benchmark when trained with Gaussian noise augmentation across multiple programming language pairs (e.g., Rust-to-Python, Java-to-Python)?.

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 6.0/10.

3 Results

11 papers retrieved. 8 claims extracted; 3 independently verified. Quality review score: 6.0/10.

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
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	×	0.10
The proposed beam search-based methodology efficiently uncovers cross-lingual weaknesses in LLMs.	✓	0.23
The code for the study is available at https://github.com/xzx34/Cross-Lingual-Pitfalls.	✓	0.26
The proficiency demonstrated in English often fails to generalize to other languages, resulting in errors in other lingu	×	0.03
The performance scores for different languages and datasets are provided in the benchmark tables.	×	0.03
The linguistic similarity scores between Chinese, Japanese, and Korean are provided in the tables.	×	0.04

References

- http://arxiv.org/abs/2310.10378v5
- http://arxiv.org/abs/2505.18673v1
- http://arxiv.org/abs/2303.12869v1