Qwen3 vs. Qwen2-1.5B Multilingual Code Generation on HumanEval-X

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does Qwen3's multilingual code generation performance compare to Qwen2-1.5B on the HumanEval-X benchmark across non-English programming languages. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CodeGeeX: A Pre-Trained Model for Code Generation with Multilingual Benchmarking on HumanEval-X. Research question: How does Qwen3's multilingual code generation performance compare to Qwen2-1.5B on the HumanEval-X benchmark across non-English programming languages?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.6/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.6/10.

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.

References

- http://arxiv.org/abs/2509.17765v1
- http://arxiv.org/abs/2303.17568v2
- http://arxiv.org/abs/2505.18673v1