Mistral-Large-2 Code Generation on MBPP: Correctness and Quality vs. Human Baselines

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do Mistral-Large-2 generated code solutions on MBPP compare to ground truth implementations in terms of functional correctness and code quality as measured by human evaluation scores. In recent years, researchers have proposed numerous benchmarks to evaluate the impressive coding capabilities of large language models (LLMs). However, current benchmarks primarily assess the accuracy of LLM-generated code, while neglecting other critical dimensions that also. 11 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: Beyond Correctness: Benchmarking Multi-dimensional Code Generation for Large Language Models. Research question: How do Mistral-Large-2 generated code solutions on MBPP compare to ground truth implementations in terms of functional correctness and code quality as measured by human evaluation scores?.

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. 11 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
Current benchmarks perform single-dimension evaluations and mostly focus only on code correctness.	×	0.09
RACE benchmark performs multi-dimensional code evaluations to identify truly high-quality code beyond correctness.	×	0.14
Code readability is the most decisive property for high-quality code.	×	0.06
Code maintainability is crucial for ensuring the software remains adaptable and easy to update, ultimately reducing long	×	0.03
Code efficiency is essential for optimizing performance, reducing resource consumption, and ensuring scalability in soft	×	0.03
Current benchmarks lack evaluation on critical dimensions that impact code quality, making it challenging to distinguish	×	0.15
Correctness-based benchmarks can lead to LLMs memorizing exact solutions from the training data instead of understanding	×	0.05
Overfitting to correctness can render LLMs susceptible to data contamination.	✓	0.16
RACE benchmark evaluates code generated by LLMs across multiple dimensions including Readability, Maintainability, Corre	✓	0.27
RACE benchmark includes metrics for Modularity, Time Efficiency, and Space Efficiency.	×	0.07
The experiment includes LLMs such as Claude-3.5-Sonnet, GPT-4o, GPT-4o-mini, GPT-3.5-turbo-0125, o1-mini-2024-09-12, Cod	×	0.01

References

- http://arxiv.org/abs/2407.11470v2
- http://arxiv.org/abs/1902.05255v1
- http://arxiv.org/abs/2505.10399v1