Mistral-Large-2 Code Correctness on MBPP: Human Evaluation Benchmark Analysis

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the human evaluation accuracy score for code correctness of Mistral-Large-2 generated solutions on the MBPP benchmark compared to reference implementations. We introduce self-invoking code generation, a new task designed to evaluate the progressive reasoning and problem-solving capabilities of LLMs. In this task, models are presented with a base problem and a related, more complex problem. 5 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: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research question: What is the human evaluation accuracy score for code correctness of Mistral-Large-2 generated solutions on the MBPP benchmark compared to reference implementations?.

2 Methodology

Systematic literature search across multiple databases yielded 15 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

15 papers retrieved. 5 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
o1-mini achieves 96.2% pass@1 on HumanEval but only 76.2% on HumanEval Pro	✓	0.28
Instruction-tuned models are less efficient on self-invoking code generation than traditional code generation tasks	✓	0.31
HumanEval and MBPP serve as fundamental benchmarks for code generation	✓	0.18
Frontier LLMs excel at generating individual code snippets	×	0.08
LLMs often struggle to effectively utilize their own generated code for solving more complex problems	×	0.08

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2211.12112v1
- http://arxiv.org/abs/2412.21199v2