Human Evaluation of Mistral-Large-2 Code Quality and Correctness on MBPP

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 score for code quality and functional correctness of Mistral-Large-2 generated solutions on MBPP compared to ground truth implementations. Several Deep Learning (DL)-based techniques have been proposed to automate code review. Still, it is unclear the extent to which these approaches can recommend quality improvements as a human reviewer. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Studying Quality Improvements Recommended via Manual and Automated Code Review. Research question: What is the human evaluation score for code quality and functional correctness of Mistral-Large-2 generated solutions on MBPP compared to ground truth 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 2.8/10.

3 Results

15 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
The study analyzed a set of 447 human reviewers' comments.	×	0.09
In the taxonomy, the number of comments labeled with a category is always higher or equal to the number of PRs containin	×	0.05
223 PRs featured at least one of the 266 comments recommending a refactoring operation.	×	0.03
Human reviewers recommended 7 rename method refactorings in the dataset.	×	0.09
ChatGPT did not suggest any of the 7 rename method refactorings recommended by human reviewers for the exact same code l	×	0.12
ChatGPT recommended the same code quality improvement as human reviewers for the exact same code location in less than 3	×	0.11
ChatGPT recommended the same code quality improvement as human reviewers for the exact same code location in more than 3	×	0.11
The default periodic compaction time was changed from 30 days to a Duration object representing 30 days.	×	0.01
A new field 'comments' with value 'comment' was added to the data structure in the code change shown in Table (p7).	×	0.02
Error handling for NoSuchFieldException and IllegalAccessException was added to log a warning about ProtobufCodec failur	×	0.01
Code was added to limit the 'RENEWED DOMAINS' environment variable to 16,000 characters and log a warning if exceeded.	×	0.02
Code was added to limit the 'FAILED DOMAINS' environment variable to 16,000 characters and log a warning if exceeded.	×	0.02

References

- http://arxiv.org/abs/1812.11470v1
- http://arxiv.org/abs/2602.11925v1
- http://arxiv.org/abs/2603.12895v1