Mistral-Large-2 Code Generation on MBPP: Automated vs. Human Evaluation Metrics

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the functional correctness and code quality of Mistral-Large-2 generated solutions on MBPP compare when evaluated using automated test suites versus human evaluation scores. The use of machine learning (ML) models to assess and score textual data has become increasingly pervasive in an array of contexts including natural language processing, information retrieval, search and recommendation, and credibility assessment of online content. A significant. 19 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: When Automated Assessment Meets Automated Content Generation: Examining Text Quality in the Era of GPTs. Research question: How does the functional correctness and code quality of Mistral-Large-2 generated solutions on MBPP compare when evaluated using automated test suites versus human evaluation scores?.

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 3.7/10.

3 Results

12 papers retrieved. 19 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
The seven ML models were each trained separately on the ASAP and CLC-FCE testbeds using 5-fold cross-validation.	×	0.08
Training each classifier within a given testbed offered better performance vis–vis training them across a consolidated	×	0.05
The dependent variable within both testbeds was standardized to a 0-1 continuous scale.	×	0.02
Two error metrics, MSE and MAE, and three agreement/correlation measures (QWK, PCC, SRC) were employed.	×	0.02
Values closer to 0 for MSE and MAE denote better performance.	×	0.01
Values closer to 1 for QWK, PCC, and SRC indicate better performance, whereas values closer to 0 signify random performa	×	0.02
BERT and RoBERTa attained the best performance across all five metrics for both datasets.	×	0.06
PRC and SRC values for BERT and RoBERTa are in the range of 0.5 to 0.76.	×	0.00
QWK results for BERT and RoBERTa are comparable to the best ASAP full dataset results attained in prior studies.	×	0.02
Direct comparisons are difficult to make because prior studies have used different problem formulations, training-testin	×	0.05
CNN and GRU outperformed feature-based methods such as SVR, XGB, and KNN.	×	0.07
Results were somewhat better on the ASAP testbed as compared to CLC-FCE for all methods on all five metrics.	×	0.03
ASAP has seen greater usage in prior AES studies.	×	0.02
The greater abundance of available training data could be the reason for better results on the ASAP testbed.	×	0.03
The benchmarking results lend credence to the ML models used.	×	0.09
ASAP dataset includes various essay types such as ARG, RESP, and NARR with different grade levels and dataset sizes.	×	0.02
FCE dataset includes various essay types such as LETT, ARG, COMM, NARR, and \$UGG with different scoring ranges.	×	0.03
ASAP 2 ARG prompt is about writing a persuasive essay reflecting views on censorship in libraries.	×	0.02
Representation examples include word, word & sense, word & POS, and word & NE formats.	×	0.01

References

- http://arxiv.org/abs/2309.14488v1
- http://arxiv.org/abs/2603.23611v1
- http://arxiv.org/abs/2204.08348v3