Multimodal Models Enhance Explainability in Low-Resource Code Generation

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of multimodal models on improving the explainability of code generation tasks in low-resource programming languages, as measured by coherence and semantic similarity scores. 14 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: Automated Collection of Evaluation Dataset for Semantic Search in Low-Resource Domain Language. Research question: What is the impact of multimodal models on improving the explainability of code generation tasks in low-resource programming languages, as measured by coherence and semantic similarity scores?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 14 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 proposed approach of using an ensemble of multiple encoders with a generative LLM (GPT-4o) to reassess relevance sco	×	0.12
The proposed approach improves the F1-score by 1.5 times compared to baseline methods.	×	0.03
Ensemble learning improves machine learning performance by combining predictions from multiple models, enhancing accurac	×	0.05
Standalone LLMs perform worse than human annotators in data annotation tasks.	×	0.05
In the implementation described, queries were generated using GPT-4o from randomly selected documents containing at leas	×	0.03
Dataset A consists of 17,053 documents, 30 queries, and 2,739 verified retrieved candidates.	×	0.02
In Dataset A, the Combined method achieved a Krippendorff's alpha of 67.03, compared to 50.30 for the Ensemble method.	×	0.04
In Dataset A, the Combined method achieved an F1-score of 53.42, while the GPT-4o-SE method achieved 46.28.	×	0.02
Dataset B consists of 14,065 documents, 30 queries, and 2,022 verified retrieved candidates.	×	0.02
In Dataset B, the Combined method achieved a Krippendorff's alpha of 68.69 and an F1-score of 49.50.	×	0.02
The average relevance score for the Combined method across categories 0 to 3 is 50.2.	×	0.04
The model 'azure-text-embedding-3-large' achieved an nDCG@10 score of 69 in the evaluation.	×	0.04
The model 'sentence-transformers/multi-qa-mpnet-base-cos-v1' achieved an average score of 35.29 across metrics P@10, R@1	×	0.04
The evaluation dataset described in Table (p11) contains 79.6K documents, 20 queries, and 406 relevant documents.	×	0.02

References

- http://arxiv.org/abs/2412.10008v1
- http://arxiv.org/abs/2605.17152v1
- http://arxiv.org/abs/2303.12869v1