Semantic Similarity Metrics Enhance Code Generation Evaluation Beyond BLEU and CodeBLEU

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of incorporating semantic similarity metrics (e.g., BERTScore) alongside BLEU and CodeBLEU on the evaluation of code generation models fine-tuned on domain-specific datasets like. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLM-as-a-Judge: Rapid Evaluation of Legal Document Recommendation for Retrieval-Augmented Generation. Research question: What is the impact of incorporating semantic similarity metrics (e.g., BERTScore) alongside BLEU and CodeBLEU on the evaluation of code generation models fine-tuned on domain-specific datasets like Android development?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

13 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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 evaluated several metrics, including Cohen's Kappa, Krippendorff's Alpha, Spearman's rank correlation, Kendall	×	0.09
The study did not evaluate every available IRR metric, such as Fleiss's Kappa and the Brennan-Prediger coefficient.	×	0.02
Bloomberg Law faces unique challenges in evaluating RAG systems due to the high accuracy standards required for legal co	×	0.05
The evaluation was conducted on two legal RAG systems operating over a comprehensive legal corpus reflecting real produc	×	0.05
Each RAG system consists of two critical components: a retrieval component that identifies relevant legal documents, and	×	0.03
System A utilizes traditional BM25 retrieval combined with an open-source LLM summarizer applied to the top 5 retrieved	×	0.04
System B incorporates improvements in the retrieval system and employs the proprietary GPT-4 model by OpenAI as the summ	×	0.02
The evaluation framework specifically targeted both the retrieval effectiveness through search relevancy assessment, and	×	0.07
Retrieval Augmented Generation (RAG) effectively enhances the capabilities of LLMs by integrating external knowledge sou	×	0.09
Traditional automated evaluation metrics such as ROUGE and BLEU depend on reference responses, which limits their effect	×	0.06
Human evaluations are typically considered the gold standard due to their accuracy, but they are impractical at large sc	×	0.02
Recent advances in LLMs have sparked interest in their potential as automated evaluators, particularly in specialized do	×	0.04
GPT-4 has shown promise in achieving human-level agreement on certain tasks.	×	0.03
Recent studies have identified key challenges in using LLMs as judges, including cognitive biases, self-preference, and	×	0.03
Researchers have explored committee-based approaches using multiple LLM and4specialized training to address limitations	×	0.05

References

- http://arxiv.org/abs/2509.12382v1
- http://arxiv.org/abs/2504.16584v1
- http://arxiv.org/abs/2509.08612v3