Semantic-Aware vs. Syntax-Only Code Models in CodeBLEU Performance Under Obfuscation

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How do semantic-aware code models compare to syntax-only models in terms of semantic consistency scores (e.g., CodeBLEU) when evaluated on the CodeXGLUE benchmark under variable name obfuscation. 16 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CodeBLEU: a Method for Automatic Evaluation of Code Synthesis. Research question: How do semantic-aware code models compare to syntax-only models in terms of semantic consistency scores (e.g., CodeBLEU) when evaluated on the CodeXGLUE benchmark under variable name obfuscation?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

16 papers retrieved. 16 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
CodeBLEU achieves a better correlation with programmer assigned scores compared with BLEU and accuracy.	✓	0.28
CodeBLEU absorbs the strength of BLEU in the n-gram match, and further injects code syntax via abstract syntax trees (AS	✓	0.37
CodeBLEU is a weighted combination of the original BLEU, the weighted n-gram match, the syntactic AST match, and the sem	×	0.11
CodeBLEU can significantly differentiate the systems' performance and achieve better correlation with the quality scores	×	0.14
BLEU measures how well a candidate translation matches a set of translation references by calculating the percentage of	×	0.03
BLEU was proposed in 2002 by Papineni et al.	×	0.03
BLEU includes a brevity penalty.	×	0.03
In the text-to-code task, System 4 (CodeGPT) has the highest CodeBLEU score of 30.96.	×	0.06
In the code translation task, System 4 (Transformer+CodeBERT) has the highest CodeBLEU score of 84.75.	×	0.05
In the code refinement task, System 3 has the highest CodeBLEU score of 83.85.	×	0.06
In the text-to-code task, System 4 (CodeGPT) has the highest human score of 3.125.	×	0.04
In the code translation task, System 4 (Transformer+CodeBERT) has the highest human score of 4.252.	×	0.04
In the code refinement task, System 3 has the highest human score of 2.022.	×	0.05
In the text-to-code task, System 4 (CodeGPT) has the highest mean CodeBLEU score of 30.13.	×	0.05
In the code translation task, System 4 (Transformer+CodeBERT) has the highest mean CodeBLEU score of 83.26.	×	0.05
In the code refinement task, System 3 has the highest mean CodeBLEU score of 82.52.	×	0.06

References

- http://arxiv.org/abs/2009.10297v2
- http://arxiv.org/abs/2102.04664v2
- http://arxiv.org/abs/2006.07180v1