Fine-Tuning JaCoText on Domain-Specific Datasets for Java Code Generation Accuracy

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of fine-tuning JaCoText on domain-specific datasets (e.g., Android development) on its accuracy for Java code generation tasks measured by BLEU and CodeBLEU scores. 16 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Out of the BLEU: how should we assess quality of the Code Generation models?. Research question: What is the impact of fine-tuning JaCoText on domain-specific datasets (e.g., Android development) on its accuracy for Java code generation tasks measured by BLEU and CodeBLEU 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 4.2/10.

3 Results

12 papers retrieved. 16 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
Human assessment is the gold standard for most machine translation or machine generation problems.	×	0.08
Manual assessment is very expensive and slow, and it is impractical to do human evaluation for each generated sample dur	×	0.07
The evaluation approaches for code generation can be split into three categories: (1) Metrics from the machine translati	×	0.14
The quality of code generation models is typically assessed by the BLEU metric score or accuracy.	×	0.14
The BLEU (BiLingual Evaluation Understudy) metric is a metric that was originally developed for the automatic quality ev	×	0.08
The BLEU metric is a corpus-level metric based on the modified n-gram precision measure with a length penalization for t	×	0.02
Researchers also consider other machine translation metrics: ROUGE-L, METEOR, and ChrF.	×	0.14
ROUGE-L is a recall-oriented metric that looks for the longest common subsequence between the reference and the candidat	×	0.02
METEOR is a mixed recall-precision metric that also penalizes candidates for not having adjacent unigrams that are adjac	×	0.01
ChrF is a character n-gram F-score metric, where precision and recall in the F-score computation are averaged over 1- to	×	0.03
Accuracy is rarely used as the primary metric in the generation tasks due to being too strict and less robust.	×	0.06
The RUBY metric was suggested by Tran et al. as an alternative to the natural languages metrics.	×	0.05
BLEU has a rather weak correlation of 0.583 with the human assessment.	×	0.05
For the HumanEval dataset, a difference in model scores of less than 5 points is considered insignificant.	×	0.12
For the HearthStone dataset, a difference in model scores of at least 2 points is enough to claim the superiority of one 4	✓	0.20
The ChrF metric is a better fit for the evaluation of code generation models than the commonly used BLEU and CodeBLEU.	✓	0.19

References

- http://arxiv.org/abs/2208.03133v2
- http://arxiv.org/abs/2509.25716v1
- http://arxiv.org/abs/2303.12869v1