To what extent does integrating domain-specific code comments from Stack Overflow improve the functional correctness of

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: To what extent does integrating domain-specific code comments from Stack Overflow improve the functional correctness of generated Java methods as measured by unit test pass rates. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: An Empirical Study of Java Code Improvements Based on Stack Overflow Answer Edits. Research question: To what extent does integrating domain-specific code comments from Stack Overflow improve the functional correctness of generated Java methods as measured by unit test pass rates?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.9/10.

3 Results

14 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 2.9/10.

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 Grid Search algorithm evaluated a total of 3,073 configurations.	×	0.03
The Grid Search process was completed in 54 hours.	×	0.02
Each configuration took an average of 63 seconds.	×	0.02
The optimized parameters resulted in the highest MRR score of 0.782.	×	0.03
Siamese never returned an empty set as results.	×	0.03
The correct clone was always present in the returned set of code snippets.	×	0.11
The correct snippet was not always displayed in the first position, occurring in only about 20% of the queries in the gr	×	0.04
An MRR value of 0.782 indicates that, in about 80% of the cases, the correct clone was ranked 1st.	×	0.03
The optimized parameters help reduce the number of false clones detected by Siamese and the time required for the manual	×	0.02
The SOTorrent dataset contains the content of 51,296,931 SO posts with 81,536,422 post versions.	×	0.02
The GHS dataset comprises 735,669 repositories written in 10 programming languages.	×	0.06
The GHS dataset provides data on 25 characteristics for each project.	×	0.04
The study searched GHS with the following filters: Language: Java; Exclude Forks; Has Open Issues; Has Open Pull Request	×	0.07
The study obtained 20,976 GitHub projects after querying GHS.	×	0.04
The statistics of the collected GitHub projects are shown in Table 4 and displayed as boxplots in Figure 5.	×	0.03
For stars, the values range from 10 to 145,397.	×	0.01
For forks, the values range from 0 to 50,497.	×	0.00
For watchers, the values range from 0 to 5,442.	×	0.00

References

- http://arxiv.org/abs/2511.05813v1
- http://arxiv.org/abs/1605.04464v1
- http://arxiv.org/abs/2412.21199v2