JaCoText Robustness to Adversarial Code Perturbations Across Python and Java

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the robustness of JaCoText to adversarial code perturbations (e.g., obfuscation, syntactic variations) vary across Python and Java when fine-tuned on different dataset sizes. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Benchmarking Robustness of Machine Reading Comprehension Models. Research question: How does the robustness of JaCoText to adversarial code perturbations (e.g., obfuscation, syntactic variations) vary across Python and Java when fine-tuned on different dataset sizes?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

4 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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 AdvRACE benchmark is constructed by applying four adversarial attacks (AddSent, CharSwap, DE, DG) on the RACE test s	×	0.09
The RACE dataset was chosen for constructing AdvRACE because it supports multiple-choice format attacks and covers diver	×	0.03
The CharSwap perturbation is applied to non-stopwords in the question and non-stopwords in the passage that have appeare	×	0.01
Each adversarial test set in AdvRACE has 4,934 examples, same as the original RACE test set.	×	0.05
Human evaluation on AdvRACE shows that all four adversarial test sets have a high valid rate close to 100% and correct r	×	0.04
AdvRACE contains four test sets for four different adversarial attacks: AddSent, CharSwap, Distractor Extraction (DE), a	×	0.07
The pipeline used to construct AdvRACE is highly efficient and transferable, allowing it to be easily adopted on other M	×	0.11
All adversarial perturbations in AdvRACE are model-agnostic and do not require access to model parameters.	×	0.07

References

- https://www.semanticscholar.org/paper/9fda15661ed7a917bb255b0ad5c0722f96f8a150
- https://arxiv.org/abs/2004.14004
- https://arxiv.org/abs/2511.19257