How does instruction tuning on secure coding guidelines affect the codeBLEU scores of Llama3-70B compared to C

Assignee Research

May 29, 2026

Abstract

Modern language models (LMs) have gained widespread acceptance in everyday and professional contexts, particularly in programming. An essential procedure enabling this adoption is instruction tuning, which substantially enhances LMs' practical utility by training them to follow user instructions and human preferences. However, existing instruction tuning schemes overlook a crucial aspect: the security of generated code. As a result, even the state-of-the-art instruction-tuned LMs frequently produce unsafe code, posing significant security risks. In this work, we introduce SafeCoder to address

1 Introduction

This paper examines: Instruction Tuning for Secure Code Generation. Research question: How does instruction tuning on secure coding guidelines affect the codeBLEU scores of Llama3-70B compared to Codestral-7B on the HumanEval-Fix task?.

2 Methodology

Systematic literature search across multiple databases yielded 8 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

8 papers retrieved. 6 claims extracted; 0 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
SafeCoder training introduces a small overhead on training time compared to standard instruction tuning due to the relat	×	0.08
SafeCoder significantly enhances the security of instruction-tuned LMs with minimal compromise on utility, e.g., Pass@1	×	0.12
State-of-the-art instruction-tuned LMs frequently produce insecure code, regardless of model size and family.	×	0.15
Even after instruction tuning, LMs still frequently produce insecure code, just like their pre-trained versions.	×	0.11
Four state-of-the-art instruction-tuned LMs generate secure code for only around 60% of the time.	×	0.14
OctoCoder, despite being tuned with general code commit data, still generates insecure code frequently.	×	0.05

References

- http://arxiv.org/abs/2312.10793v3
- http://arxiv.org/abs/2506.11022v2
- http://arxiv.org/abs/2402.09497v2