Fine-Tuning GLM-4.5 on Self-Invoking Code Datasets Enhances MBPP Pro Robustness

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does fine-tuning on self-invoking code datasets improve GLM-4.5's robustness against distribution shifts in the MBPP Pro benchmark compared to base Codex models. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research question: Does fine-tuning on self-invoking code datasets improve GLM-4.5's robustness against distribution shifts in the MBPP Pro benchmark compared to base Codex models?.

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 3.8/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

References

- http://arxiv.org/abs/2604.25392v1
- http://arxiv.org/abs/2112.03057v1
- http://arxiv.org/abs/2412.21199v2