CodeLlama-7B Reasoning Accuracy on BinMetric vs. Specialized Binary Analysis LLMs

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: What is the comparative reasoning accuracy of codellama-7b-hf-float16 on binary analysis tasks versus other specialized LLMs (e.g., BinGPT, assemblyLLM) using the BinMetric benchmark. 10 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can LLMs Compress (and Decompress)? Evaluating Code Understanding and Execution via Invertibility. Research question: What is the comparative reasoning accuracy of codellama-7b-hf-float16 on binary analysis tasks versus other specialized LLMs (e.g., BinGPT, assemblyLLM) using the BinMetric benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.8/10.

3 Results

1 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 8.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
LLMs demonstrate strong performance on code benchmarks.	✓	0.23
Consistent reasoning across forward and backward execution remains elusive for LLMs.	✓	0.26
RoundTripCodeEval (RTCE) is a benchmark of four code execution reasoning tasks.	✓	0.26
RTCE evaluates round-trip consistency through execution-free, exact-match assessment of bijection fidelity across four l	✓	0.35
State-of-the-art Code-LLMs were evaluated under zero-shot prompting, supervised fine-tuning on execution traces, and ite	✓	0.33
All approaches yield only modest improvements and none closes the gap, revealing that current LLMs lack the internal coh	✓	0.37
RTCE surfaces findings invisible to existing benchmarks: models frequently pass individual forward and backward tasks ye	✓	0.44
SFT and self-reflection saturate after one revision round, indicating they cannot repair fundamental algorithmic misunde	✓	0.31
Failures persist even on simple bijections such as RLE, suggesting that algorithmic complexity is not the sole root caus	✓	0.29
Code and dataset are available at https://github.com/Nickil21/round-trip-code-compression.	✓	0.31

References

- `https://openalex.org/W7125353145`