SpikingBrain and Claude 3 Haiku Accuracy-Latency Trade-offs on MBPP-X

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the tradeoff between pass@1 accuracy and latency (ms/token) when comparing SpikingBrain's multi-agent pipeline with Claude 3 Haiku's single-model approach on the MBPP-X benchmark. 10 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can Open Large Language Models Catch Vulnerabilities?. Research question: What is the tradeoff between pass@1 accuracy and latency (ms/token) when comparing SpikingBrain's multi-agent pipeline with Claude 3 Haiku's single-model approach on the MBPP-X benchmark?.

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 6.9/10.

3 Results

4 papers retrieved. 10 claims extracted; 6 independently verified. Quality review score: 6.9/10.

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 study evaluates three LLMs: Llama3, Codestral, and Deepseek R1.	×	0.15
The evaluation uses a carefully filtered subset of the Big-Vul dataset.	✓	0.17
The dataset subset is annotated with eight representative Common Weakness Enumeration (CWE) categories.	✓	0.17
The study adopts a closed-world classification setup.	×	0.12
The evaluated models demonstrate high detection rates for the presence of vulnerabilities.	×	0.14
The evaluated models demonstrate markedly poor accuracy in mapping vulnerabilities to the correct CWE label.	✓	0.15
The models exhibit frequent overgeneralization and misclassification when assigning CWE labels.	×	0.09
The study analyzes model-specific biases and common failure modes.	✓	0.17
Current LLMs have limitations in performing fine-grained security reasoning.	✓	0.21
LLMs are being adopted as learning aids in educational contexts despite their limitations.	✓	0.20

References

- https://doi.org/10.4230/oasics.icpec.2025.4
- https://openalex.org/W7124228147
- https://doi.org/10.3390/s20092533