Claude-3.5-Haiku Benchmark Performance Across Reasoning and Coding Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Claude-3.5-Haiku on reasoning mathematics coding and language understanding tasks. 9 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Agentless: Demystifying LLM-based Software Engineering Agents. Research question: What are the benchmark performance scores of Claude-3.5-Haiku on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.8/10.

3 Results

11 papers retrieved. 9 claims extracted; 5 independently verified. Quality review score: 6.8/10.

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
Agentless employs a three-phase process consisting of localization, repair, and patch validation.	✓	0.18
Agentless does not let the LLM decide future actions or operate with complex tools.	✓	0.21
Agentless achieved a performance score of 32.00% on the SWE-bench Lite benchmark.	×	0.12
Agentless produced 96 correct fixes on the SWE-bench Lite benchmark.	✓	0.15
Agentless incurred a cost of $0.70 on the SWE-bench Lite benchmark.	×	0.13
Agentless achieved the highest performance compared with all existing open-source software agents on the SWE-bench Lite	✓	0.26
The authors manually classified problems in the SWE-bench Lite benchmark.	✓	0.17
The manual classification of SWE-bench Lite identified problems containing exact ground truth patches.	×	0.11
The manual classification of SWE-bench Lite identified problems with insufficient or misleading issue descriptions.	×	0.15

References

- https://doi.org/10.48550/arxiv.2407.01489
- https://doi.org/10.48550/arxiv.2501.18362
- https://doi.org/10.48550/arxiv.2411.04872