GPT-4o SWE-Bench Score Discrepancies Across Evaluation Protocols

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: Benchmark archaeology: investigate SWE-bench score discrepancy for GPT-4o — reported 7.0%–83.4% (spread 76.4pp) across 2 papers. Sources: 'SWE-bench Goes Live!' (7.0%); 'FeedbackEval: A Benchmark for. The issue-resolving task, where a model generates patches to fix real-world bugs, has emerged as a critical benchmark for evaluating the capabilities of large language models (LLMs). While SWE-bench and its variants have become standard in this domain, they suffer from key. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: SWE-bench Goes Live!. Research question: Benchmark archaeology: investigate SWE-bench score discrepancy for GPT-4o — reported 7.0%–83.4% (spread 76.4pp) across 2 papers. Sources: 'SWE-bench Goes Live!' (7.0%); 'FeedbackEval: A Benchmark for Evaluating' (83.4%). Identify evaluation protocol differences (few-shot, prompting, preprocessing)..

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

9 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
SWE-bench-Live was evaluated using three agent frameworks: OpenHands, SWE-Agent, and Agentless.	×	0.13
OpenHands was set to a maximum of 60 iterations per instance.	×	0.04
SWE-Agent was limited to 100 LLM calls per instance.	×	0.04
Agentless was evaluated without the reranking stage based on regression testing.	×	0.03
Four LLMs were used for testing: GPT-4o, GPT-4.1, Claude 3.7 Sonnet, and DeepSeek V3.	×	0.07
The primary evaluation metric is Resolved Rate (%).	×	0.02
Patch Apply Rate (%) measures the percentage of syntactically correct patches.	×	0.03
Localization Success Rate (%) reflects whether the modified files match the gold patch.	×	0.01
The highest resolved rate on SWE-bench-Live is 19.25%.	×	0.10
Recent state-of-the-art agents and models report a resolved rate exceeding 60% on the SWE-bench Verified subset.	×	0.13
OpenHands with GPT-4o achieved a resolved rate of 7.00%, a patch apply rate of 72.00%, and a localization success rate o	×	0.05
SWE-Agent with GPT-4o achieved a resolved rate of 10.00%, a patch apply rate of 93.33%, and a localization success rate	×	0.06
Agentless with GPT-4o achieved a resolved rate of 11.67%, a patch apply rate of 91.67%, and a localization success rate	×	0.04

References

- http://arxiv.org/abs/2603.00520v1
- http://arxiv.org/abs/2308.10783v2
- http://arxiv.org/abs/2505.23419v2