SWE-bench Verified

SWE-BENCH-VER · agentic benchmark · 2024

Source: https://policywindow.org/wiki/swe-bench-verified

Generated 2026-05-30T22:10:25 UTC

Summary

Solve real-world GitHub issues from 12 popular Python repos. The 'Verified' subset is human-validated to remove ambiguity and have working tests.

At a glance

Score range

0–100 % solved

Contamination risk

medium

Methodology

https://openai.com/index/introducing-swe-bench-verified/

Saturation

active

Details

500-task verified subset. Run-time evaluation; can't be gamed by pure memorisation but agent harness affects results.

How to cite this article

APA

Policy Window. (2024). SWE-bench Verified [Wiki article — Benchmark]. https://policywindow.org/wiki/swe-bench-verified

CHICAGO

Policy Window. 2024. "SWE-bench Verified." Wiki article (Benchmark). https://policywindow.org/wiki/swe-bench-verified.

HARVARD

Policy Window (2024) 'SWE-bench Verified', Wiki article — Benchmark, available at: https://policywindow.org/wiki/swe-bench-verified.

OSCOLA

Policy Window, 'SWE-bench Verified' (Wiki article — Benchmark, 2024) <https://policywindow.org/wiki/swe-bench-verified> accessed [date].

BIBTEX

```
@misc{policywindow-swe-bench-verified,
title  = {SWE-bench Verified},
author = {Policy Window},
year   = {2024},
howpublished = {SWE-BENCH-VER (2024)},
url    = {https://policywindow.org/wiki/swe-bench-verified},
```

```
note   = {Primary source: https://openai.com/index/introducing-swe-bench-verified/}
}
```