SWE-rebench V2: A 20-Language Benchmark for AI Coding Agents

Kabui, Charles

2026-03-13

Read at ToKnow.ai

Nebius released SWE-rebench V2, a dataset of 32,079 real software engineering tasks spanning 20 programming languages and 3,600+ open-source repositories. The original SWE-bench, which every major AI lab uses to report coding agent scores, only covers Python. SWE-rebench V2 adds Go, TypeScript, JavaScript, Rust, Java, C, C++, C#, PHP, Kotlin, Julia, Elixir, Scala, Swift, Dart, R, Clojure, OCaml, and Lua. Each task comes from a real GitHub

pull request, with fail-to-pass tests and pre-built Docker images for reproducible execution. An automated pipeline handles the collection: an interactive setup agent synthesizes installation and test procedures per repository, and an ensemble of LLM judges filters out unreliable instances. Nebius also releases 120,000+ additional tasks with metadata for scaling up reinforcement learning training.

AI coding agents are one of the fastest-moving areas in AI, and every major model reports SWE-bench scores. But Python-only evaluation hides real differences in how models handle statically typed languages, compiled languages, and different toolchains. A model scoring 72% on Python tasks might struggle with Rust's borrow checker or Go's concurrency patterns. SWE-rebench V2 lets researchers test across the languages developers actually use, with 3,600+ repositories representing real codebases, not synthetic exercises. Qwen3.5 scored 69.2% on SWE-bench Verified, but we don't yet know how it handles TypeScript or Rust at scale.

The shift from single-language benchmarks to multilingual ones mirrors what happened in NLP years ago. If coding agents are going to replace real engineering work, they need to be tested on real engineering diversity.

Sources:

- SWE-rebench V2 Paper (arXiv)
- SWE-rebench V2 Dataset on Hugging Face
- SWE-rebench V2 GitHub Repository
- SWE-bench Verified Dataset
