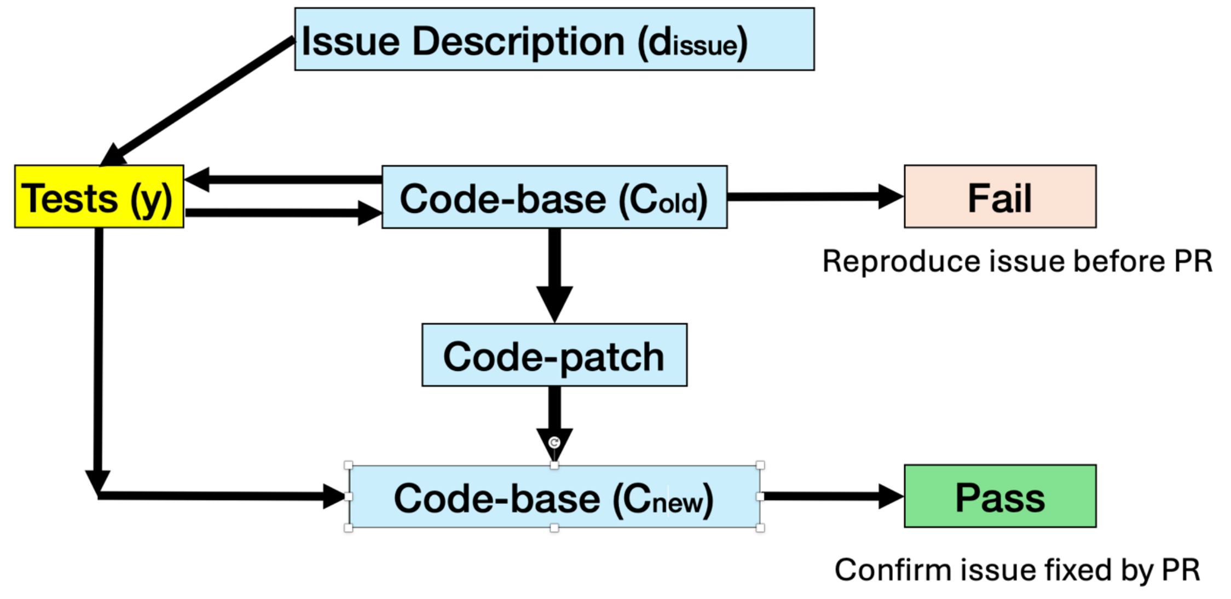
Otter: Generating Tests from Issues to Validate SWE Patches

Toufique Ahmed, Jatin Ganhotra, Rangeet Pan, Avraham Shinnar, Saurabh Sinha, Martin Hirzel

IBM Research

Problem Statement

Let d_{issue} = issue description, c_{old} = old code before PR, and c_{new} = new code after PR. The problem is to generate tests y given as input only x = (d_{issue}, c_{old}), without access to c_{new}.


Issue Description (dissue)
Tests (y)
Code-base (Cold)
Fail
Reproduce issue before PR
Code-patch
Code-base (Cnew)
Pass
Confirm issue fixed by PR

Contributions

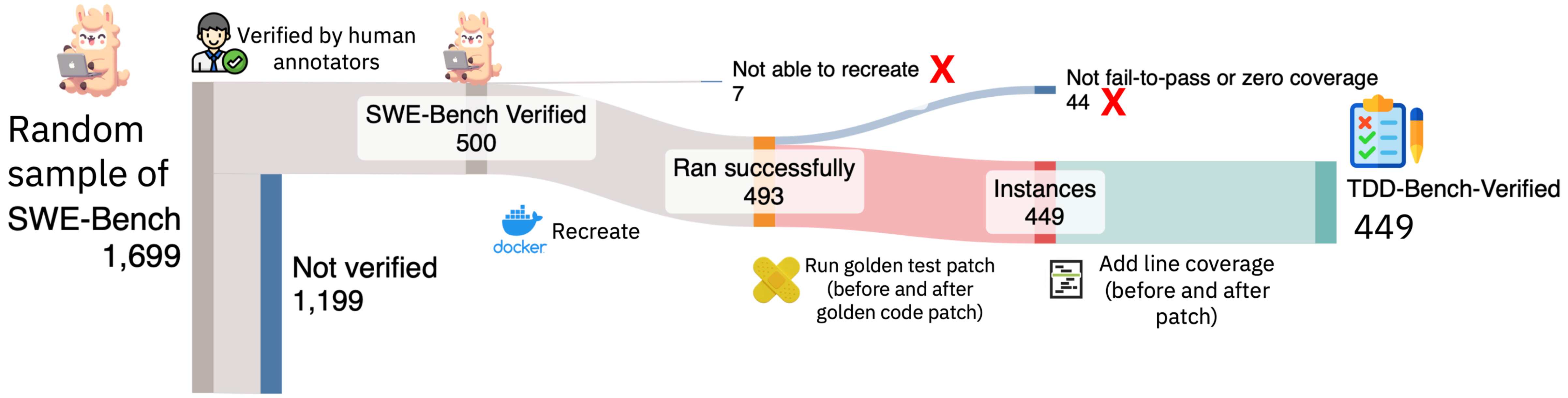
- Two bug reproduction test generation approaches: i) Otter ii) Otter++
- Benchmark to evaluate reproduction tests: TDD-Bench-Verified.

Motivation

- To systemically evaluate test generation tools (using Benchmark)
- Improve precision of SWE-agents by validating SWE-patches
- Support Test Driven Development (TDD)
  - make requirement more precise
  - easy to maintain codebase

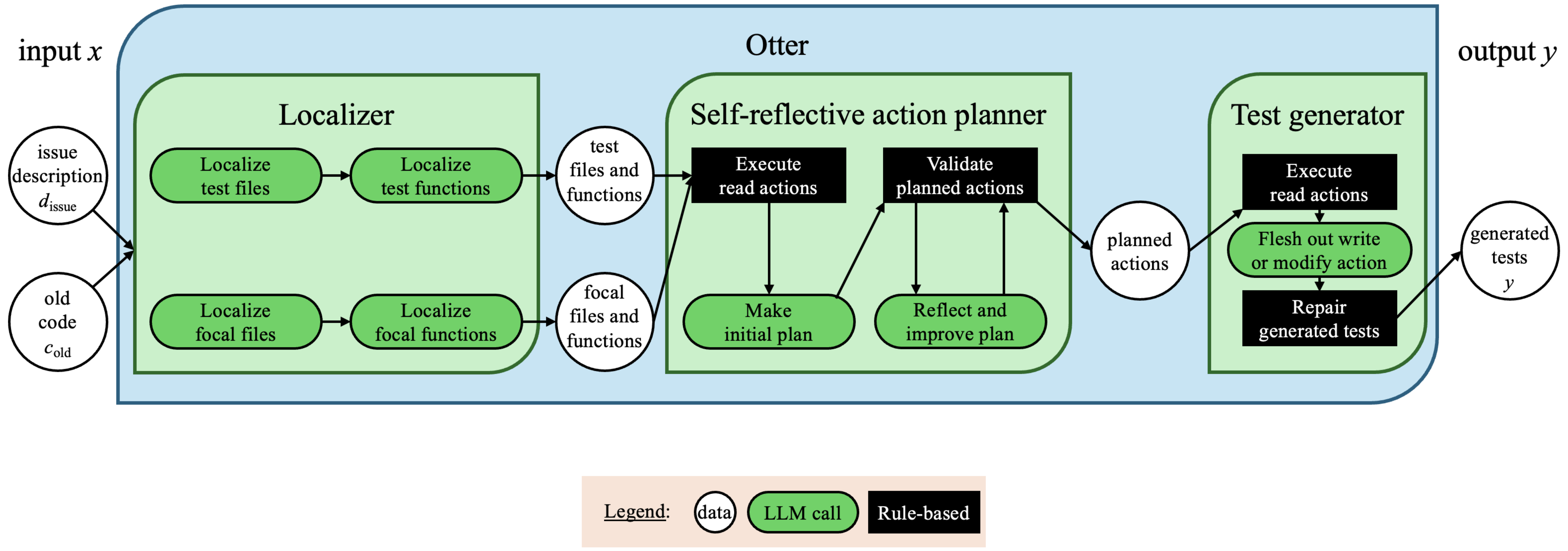
TDD-Bench-Verified: ***https://github.com/IBM/TDD-Bench-Verified***

Benchmark Construction


Random sample of SWE-Bench
1,699
Verified by human annotators
SWE-Bench Verified
500
Not verified
1,199
Recreate
docker
Not able to recreate
7
Ran successfully
493
Not fail-to-pass or zero coverage
44
Run golden test patch (before and after golden code patch)
Instances
449
Add line coverage (before and after patch)
TDD-Bench-Verified
449

- Started with 500 samples from SWE-Bench-Verified
- Ended up with 449 after all filtering process (based on coverage and f2p property)
- We propose a new metric tddScore (consider coverage also)

Otter: Overview

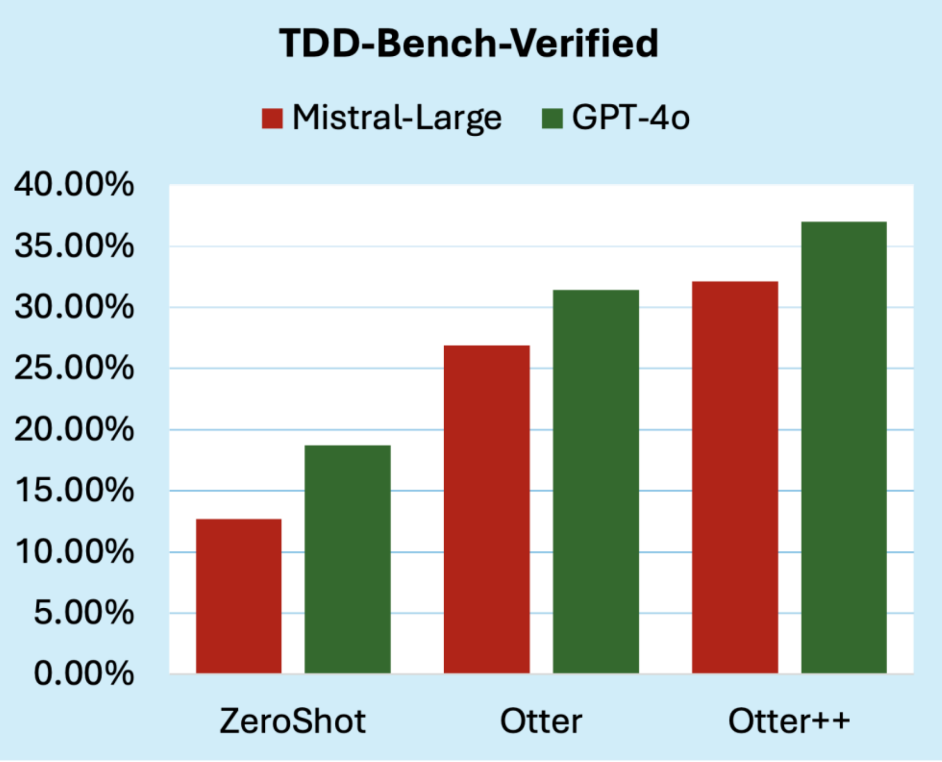


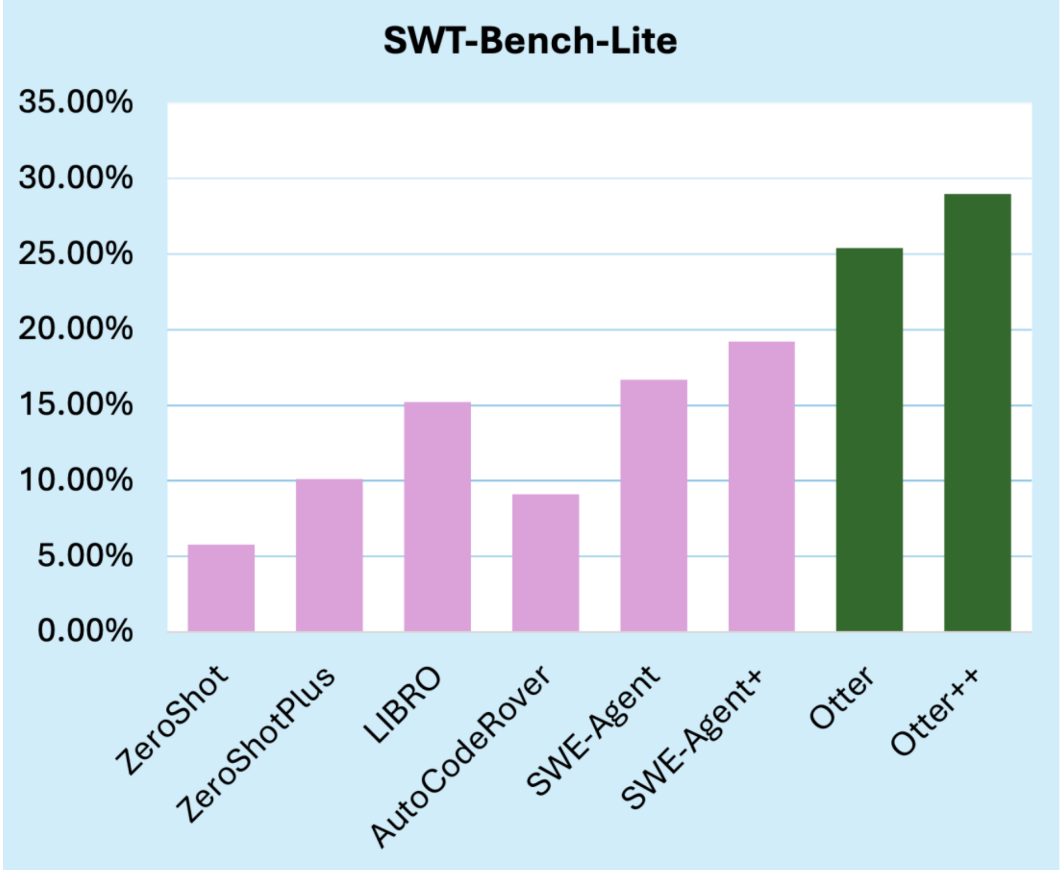
We have three components:

i. Localizer
ii. Self-reflective action planner
iii. Test generator

Performance Evaluation

- Otter++ is Otter with inference scaling

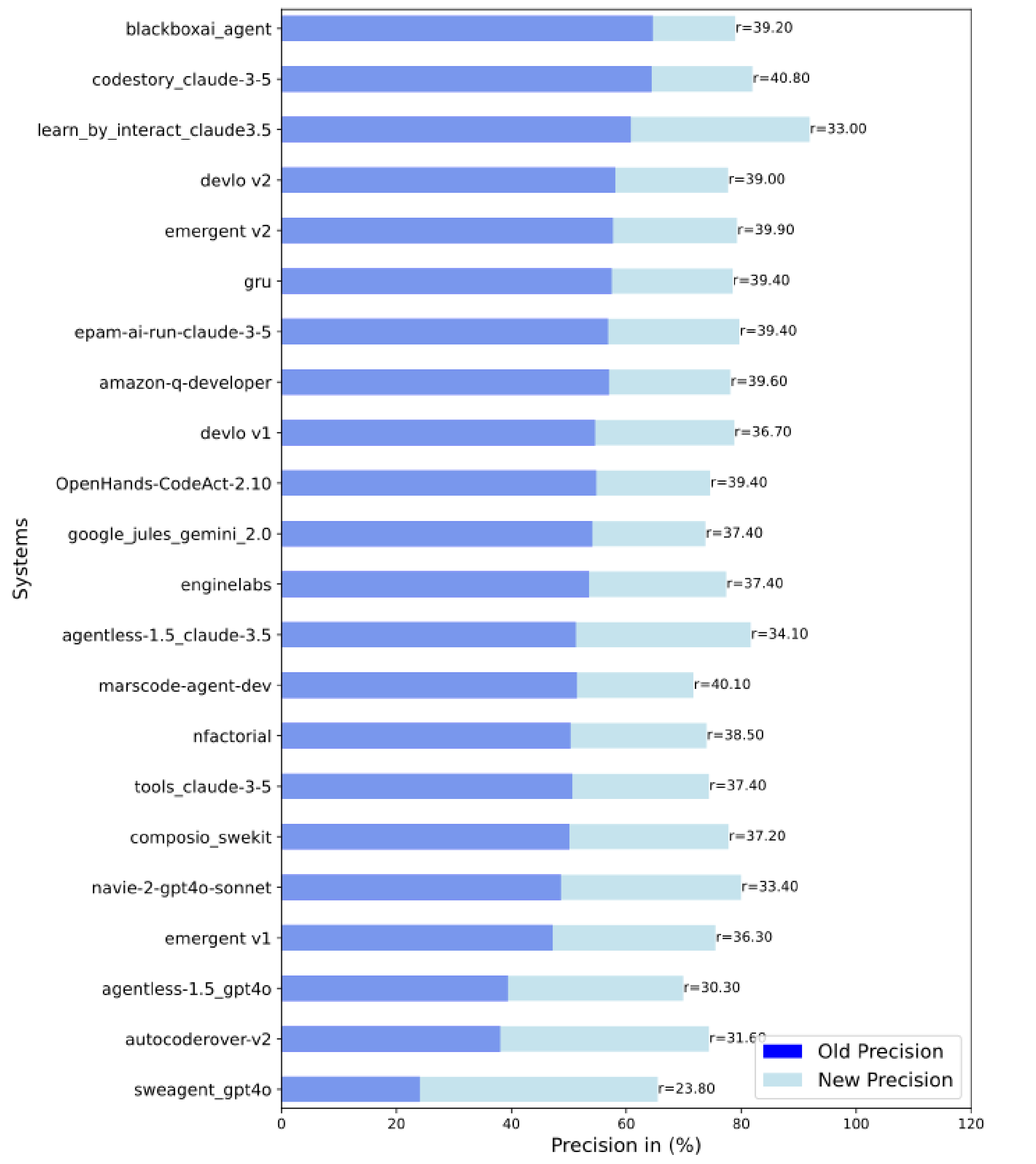




Key Insights

- Without self reflective action planning, we lose more than 14%-20% of f2p tests for GPT-4o and 21%-36% for Mistral-large model (see details in the paper).
- For each instance, Otter costs $0.06 and Otter++ costs $0.09

Validating SWE-Patches



- 65% to 92% precision while maintaining a decent recall of 30%-41%.

Conclusion

- Proposed Otter, a system that generates tests from issues, using LLMs with a novel self-reflective action planner
- Open-sourced TDD-Bench-Verified, a benchmark for test driven development
- An empirical study on using tests generated from issues to filter SWE-Patches

Forty-Second International Conference on Machine Learning, Vancouver, Canada, July 13-19, 2025