Otter: Generating Tests from Issues to Validate SWE Patches

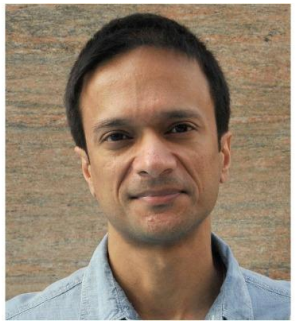
IBM Research

Toufique Ahmed

Jatin Ganhotra

Rangeet Pan

Avraham Shinnar



Saurabh Sinha

Martin Hirzel

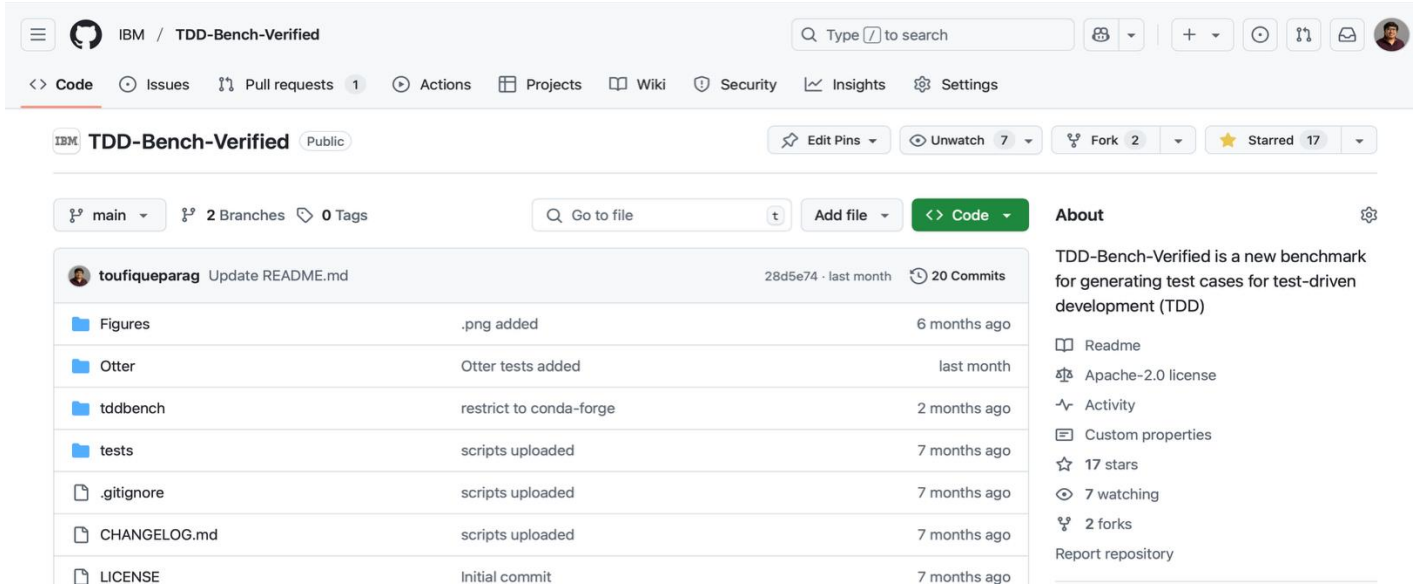
Research

IBM

Contributions

- Two bug reproduction test generation approaches: i) Otter ii) Otter++

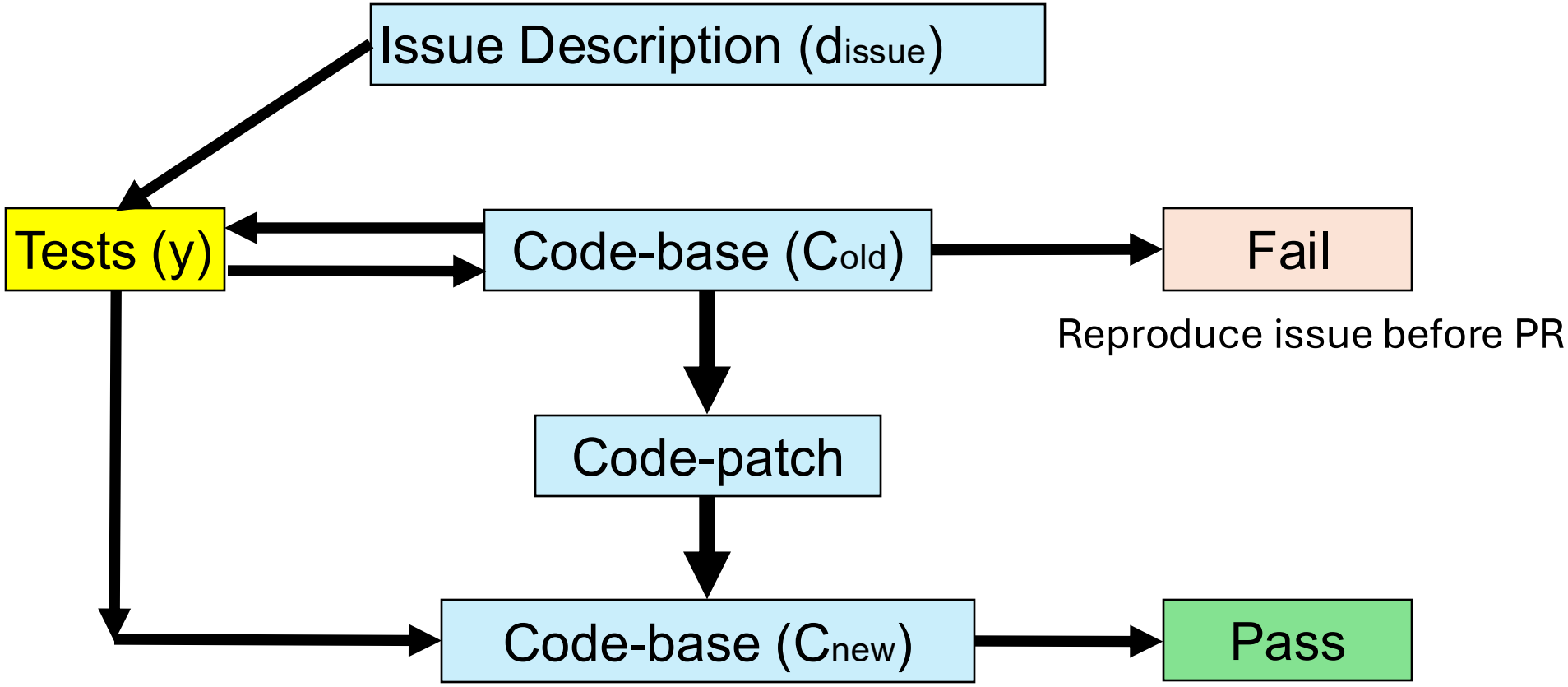
- Benchmark to evaluate reproduction tests: TDD-Bench-Verified



Research

IBM®

Problem Statement



Research

IBM

Motivation

- To systemically evaluate test generation tools (using Benchmark)

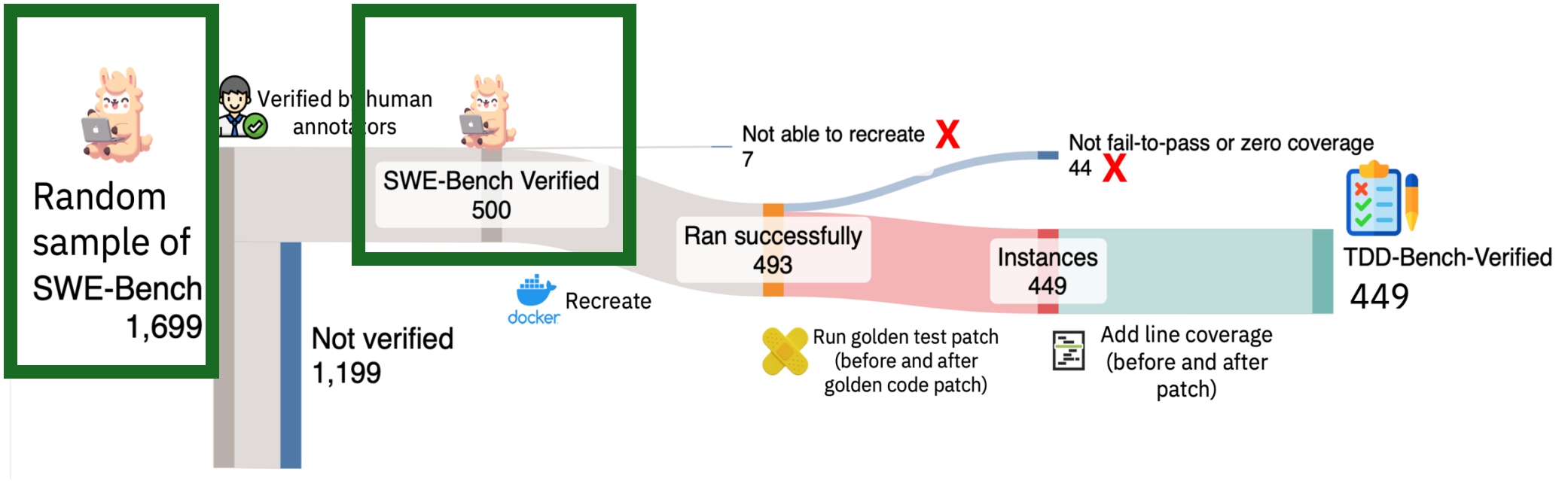
- Improve precision of SWE-agents by validating SWE-patches

- To Support Test-driven Development (TDD)
  - Make requirement more precise
  - Easy to maintain code-base

Research

IBM®

Constructing the Benchmark

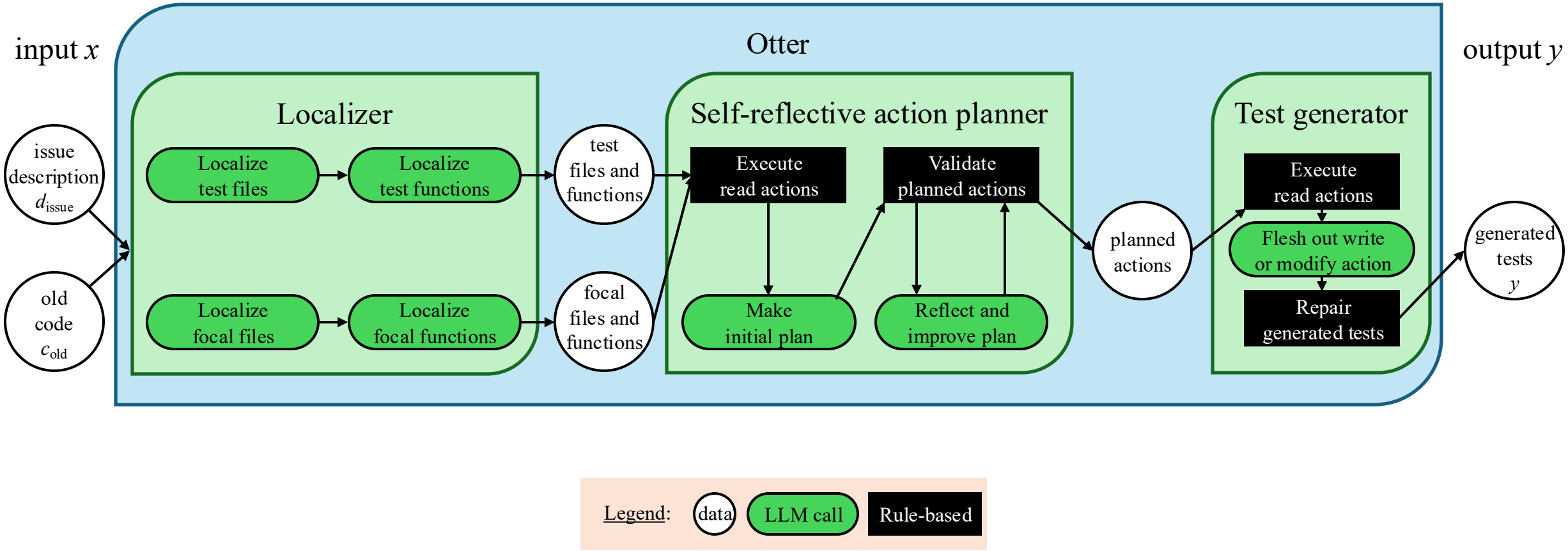


- Started with 500 samples proposed by OpenAI
- Ended up with 449 after all filtering process
- We propose a new metric tddScore (consider coverage also)

Research

IBM

Otter: Overview

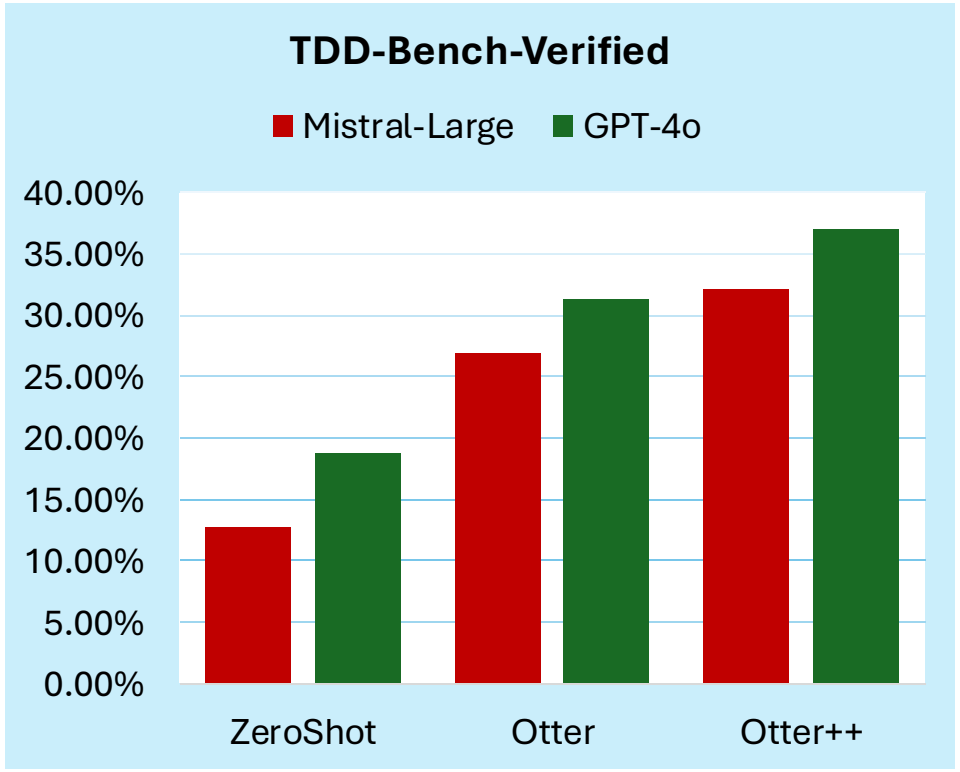


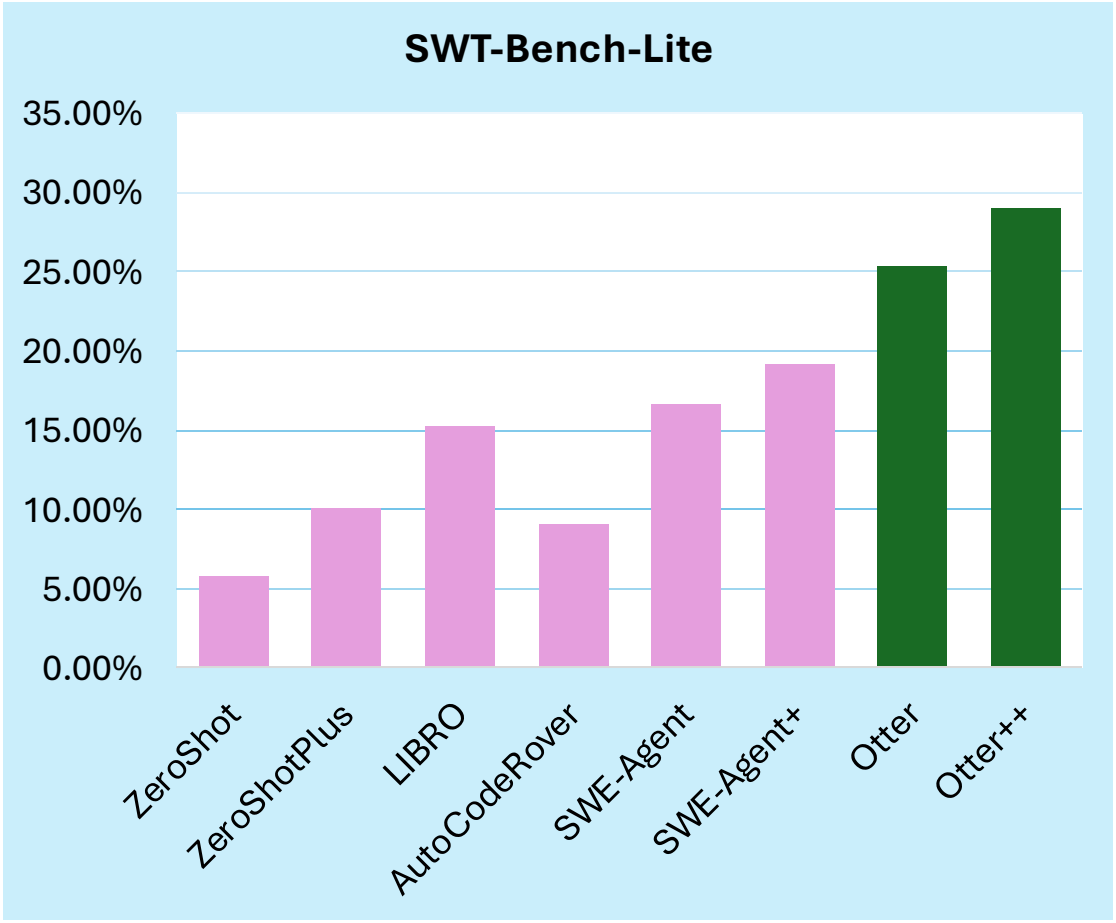
Research

IBM

Comparing with Other Approaches

Otter++ is Otter with inference scaling

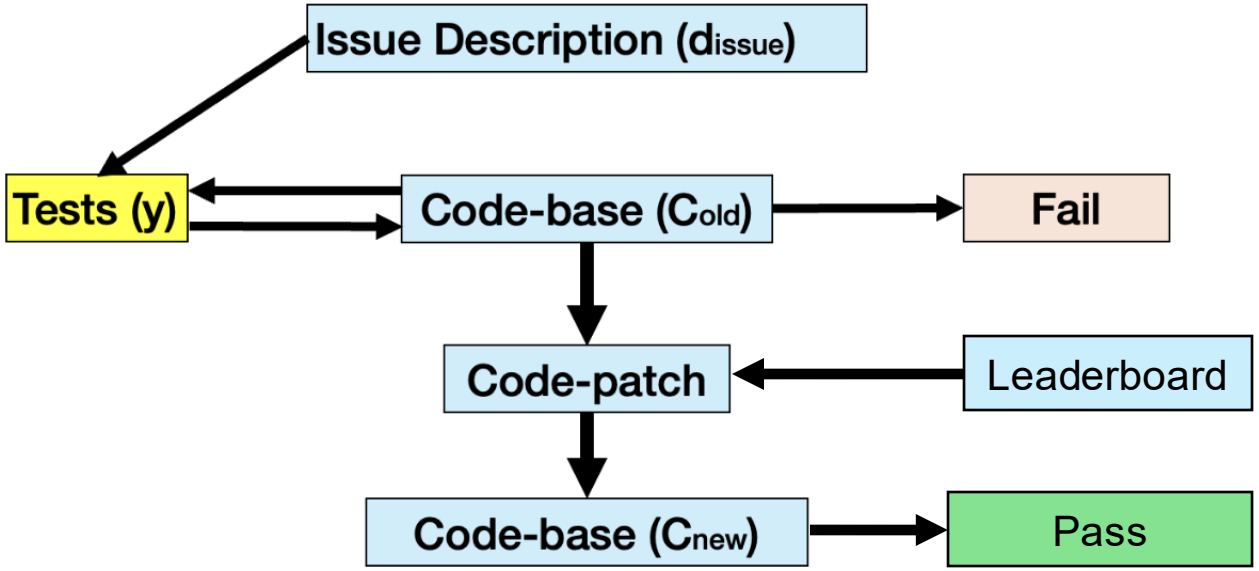


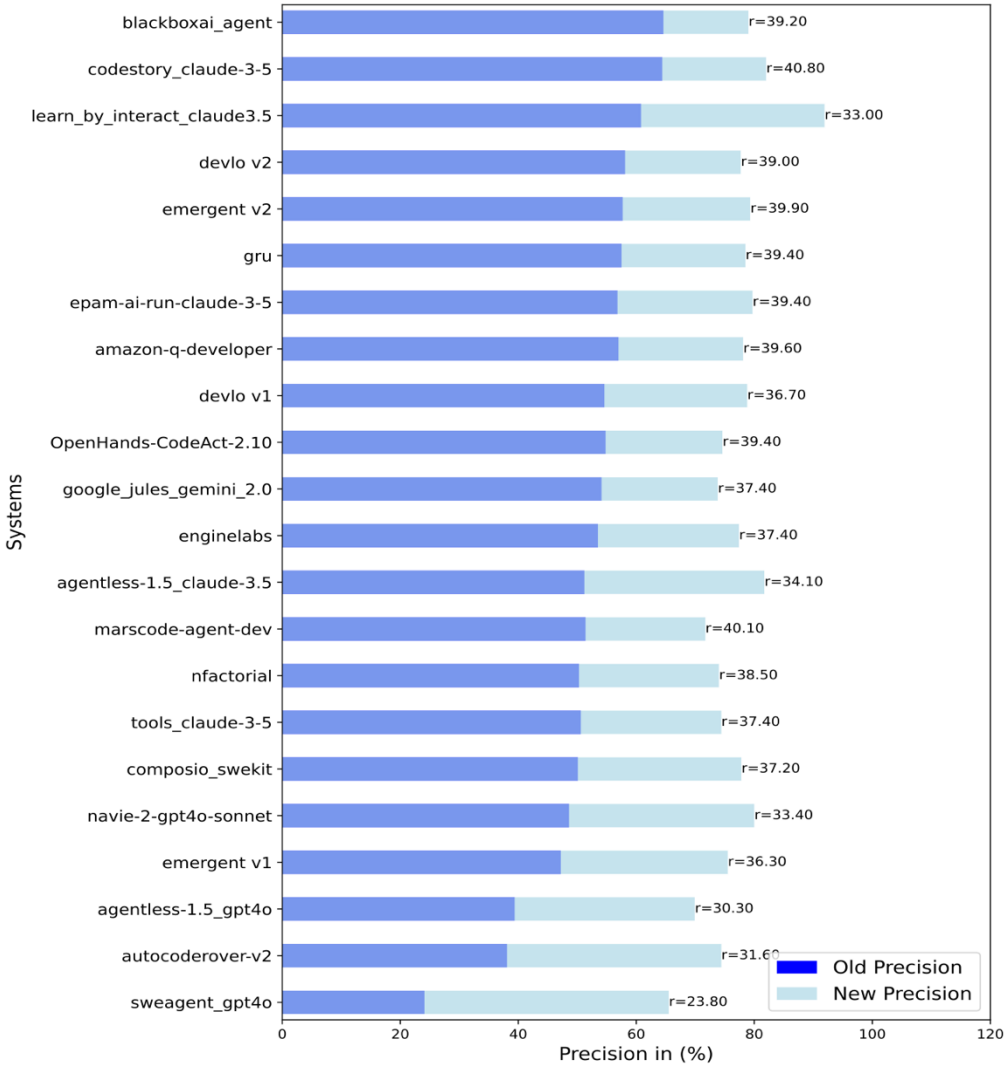


Research

IBM

Validating SWE-Patches using Our Tests





- Ran 5 sets of tests on 22 system from SWE-bench-verified.
- Only submit patch if fail-to-pass

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

Research

IBM®

Conclusion

- Otter, a system that generates tests from issues, using LLMs with a novel self-reflective action planning.
- TDD-Bench-Verified, a benchmark for test driven development with a new metric tddScore
- An empirical study on using tests generated from issues to filter issue-resolution candidates for SWE-bench Verified.

Research

IBM