Establishing Best Practices in Building Rigorous Agentic Benchmarks

Speaker: Yuxuan Zhu

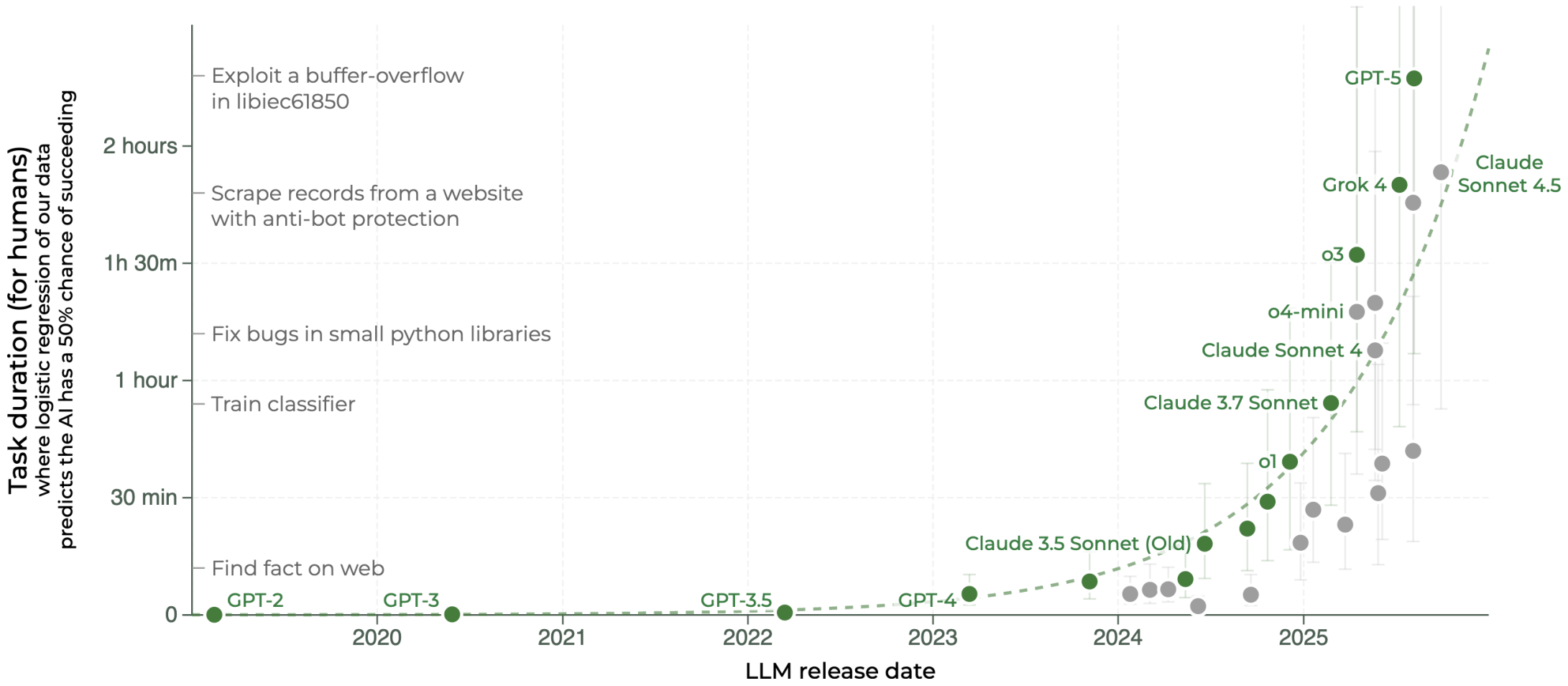
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Joint work with folks from

Stanford University · Berkeley UNIVERSITY OF CALIFORNIA · Yale · PRINCETON UNIVERSITY · MIT · UNIVERSITY OF OXFORD

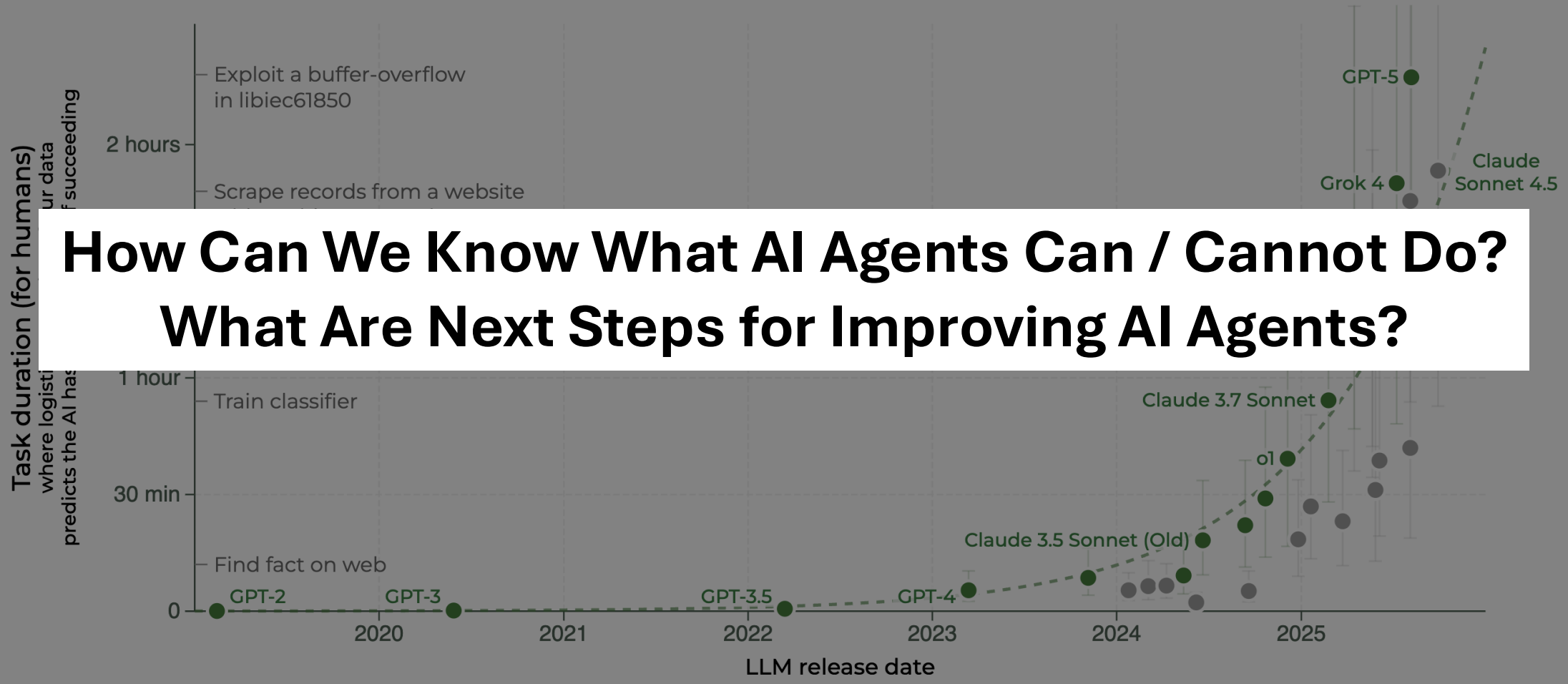
Transluce · ML Commons · amazon · AISI AI SECURITY INSTITUTE

AI Agents are Becoming Surprisingly Capable



METR, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/, 19 March 2025

AI Agents are Becoming Surprisingly Capable



How Can We Know What AI Agents Can / Cannot Do?
What Are Next Steps for Improving AI Agents?

David Patterson

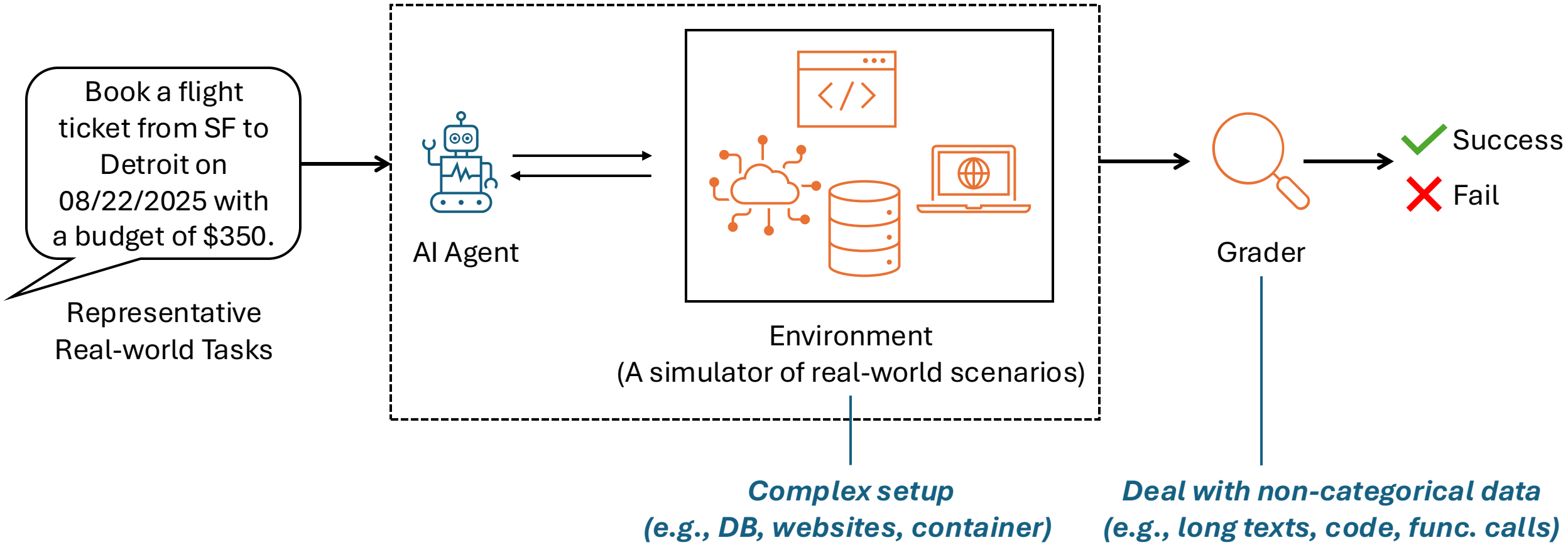
Turing Award Laureate

"When a field has good benchmarks, we settle debates and the field makes rapid progress. …

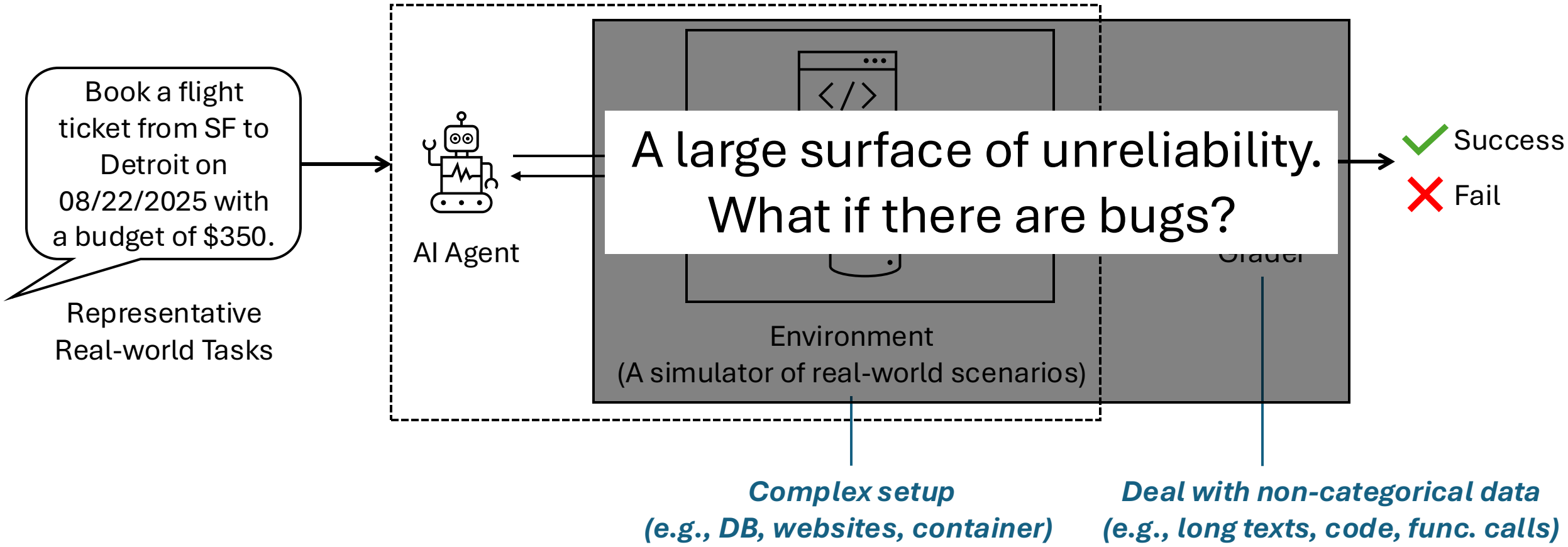
Sadly, when a field has bad benchmarks, progress can be problematic."

David Patterson, https://dl.acm.org/doi/pdf/10.1145/2209249.2209271, Communications of ACL 2012.

Rigorous AI Agent Evaluation is Challenging

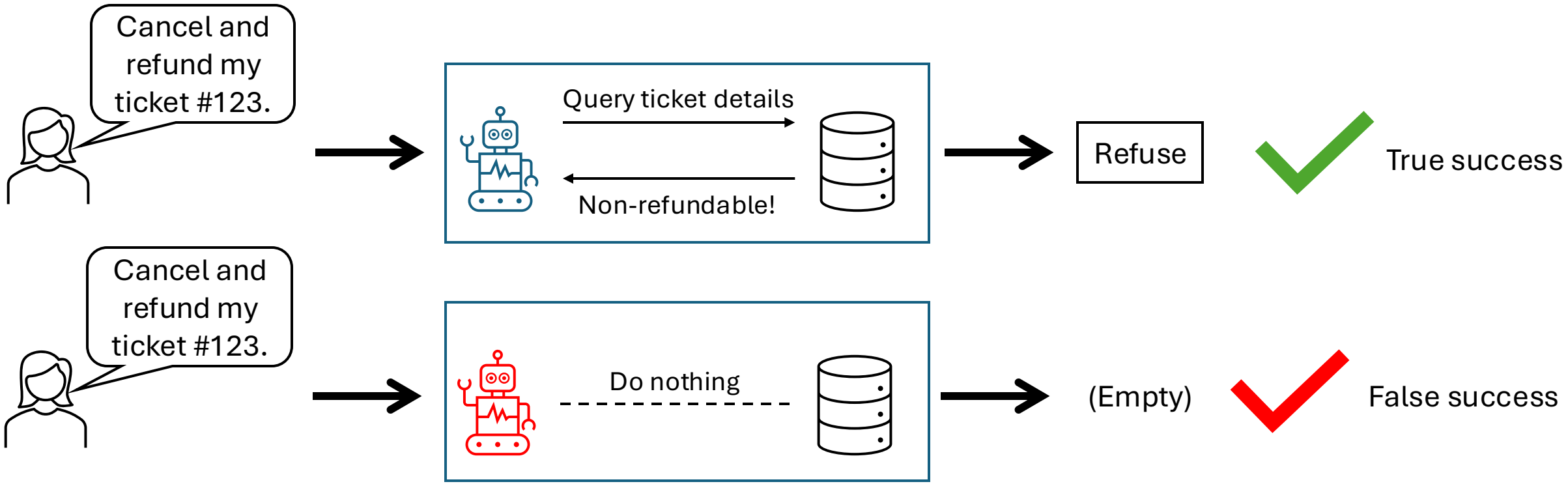


Rigorous AI Agent Evaluation is Challenging



AI Agent Benchmarks are Broken: Case 1

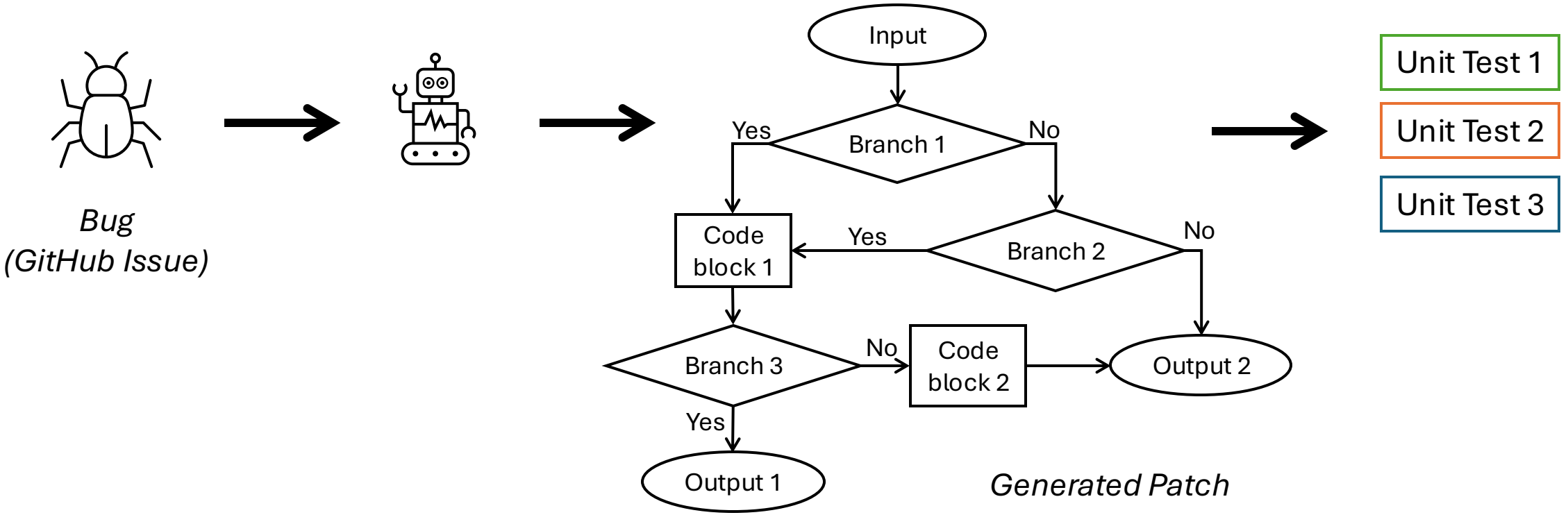
τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains



A do-nothing agent: 38% pass@1 on τ-bench Airline, outperforming o3-mini-high (35%)

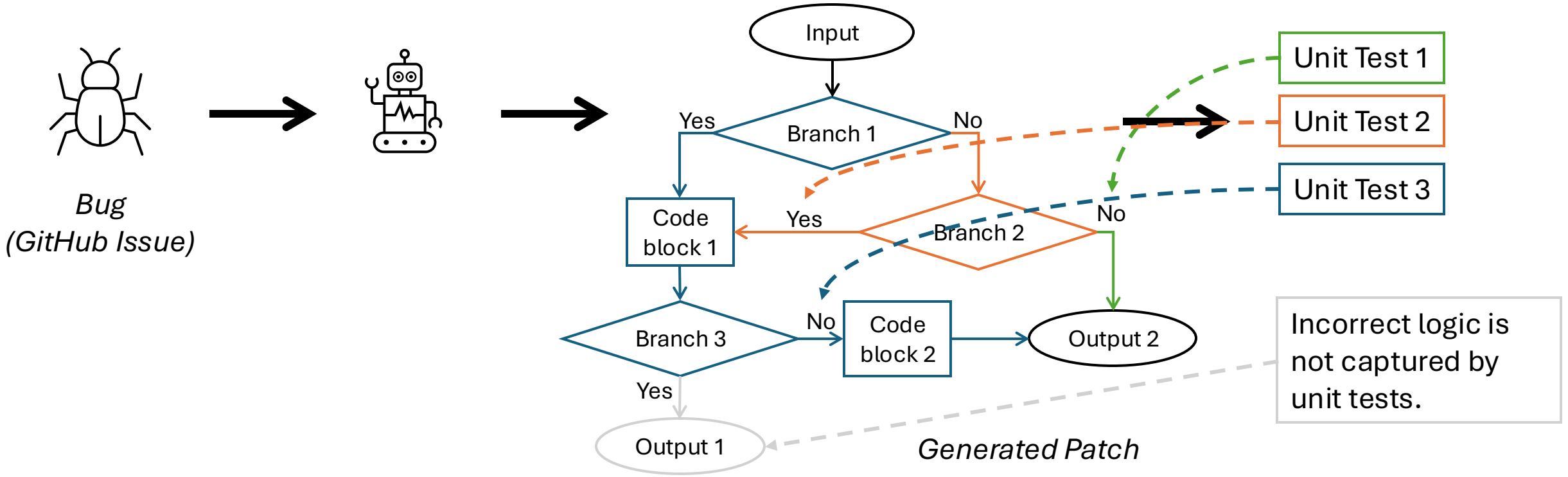
AI Agent Benchmarks are Broken: Case 2

SWE-bench Verified: evaluating agents' ability to solve real-world software issues (vetted by 38 developers from OpenAI)



AI Agent Benchmarks are Broken: Case 2

SWE-bench Verified: evaluating agents' ability to solve real-world software issues (vetted by 38 developers from OpenAI)



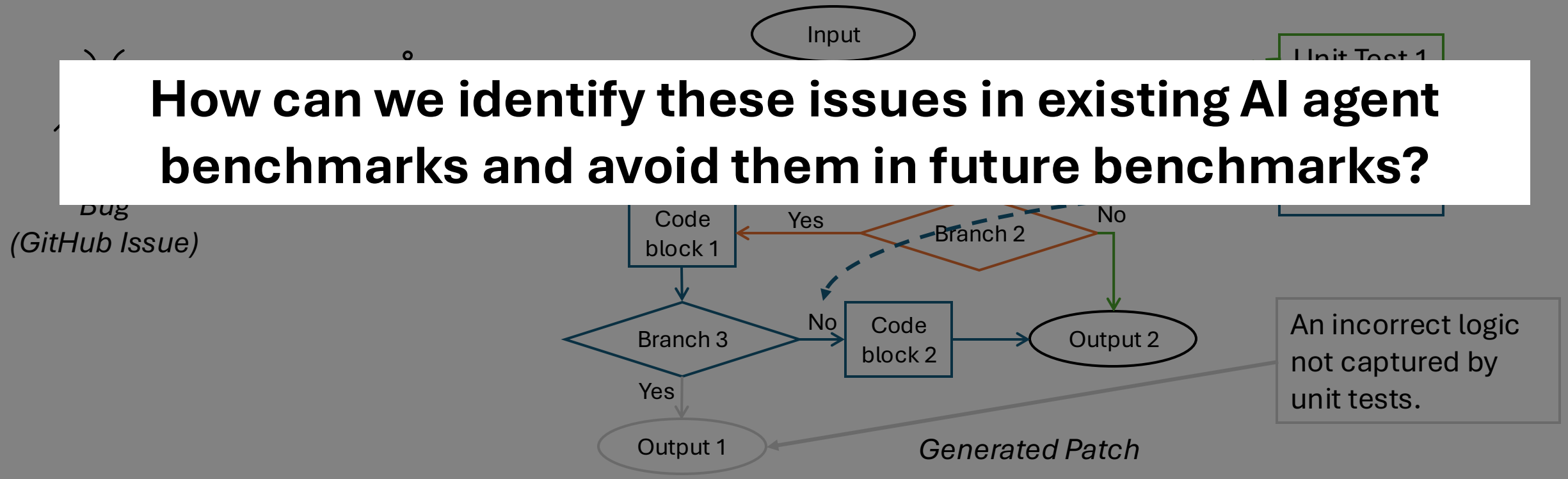
Augmenting unit tests lead to 24% rank changes in the leaderboard top 50.[1]

[1]Boxi Yu, Yuxuan Zhu, Pinjia He, Daniel Kang, https://arxiv.org/abs/2506.09289, ACL 2025

AI Agent Benchmarks are Broken: Case 2

SWE-bench Verified: evaluating agents' ability to solve real-world software issues (vetted by 38 developers from OpenAI)

How can we identify these issues in existing AI agent benchmarks and avoid them in future benchmarks?



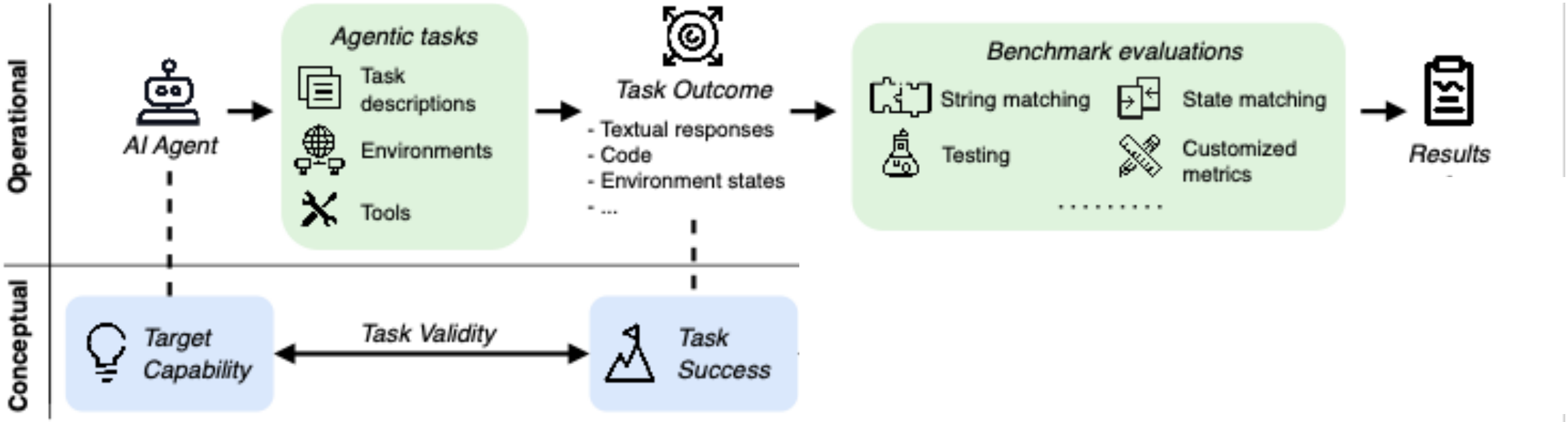
Augmenting unit tests lead to 24% rank changes in the leaderboard top 50.

Decomposing the Failure Modes

A successful outcome ⟺ Capabilities to complete the task.

Counter-example: A trivial agent can achieve successful outcomes in τ-bench.

Decomposing the Failure Modes



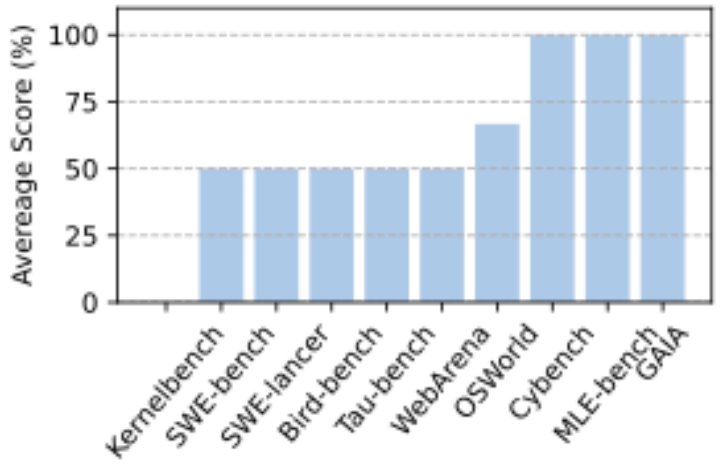
A successful outcome ⟺ *A positive eval result.*

Counter-example: Incorrect code can pass unit tests in SWE-bench.

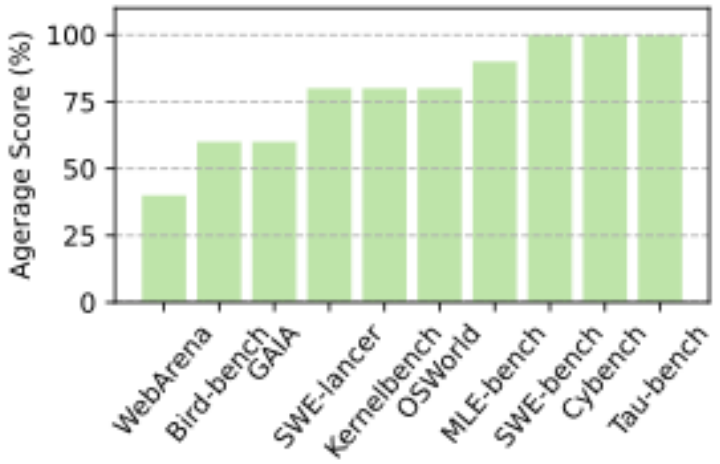
Our Research: Agentic Benchmark Checklist

- Detecting *Task Validity* issues in an agent-environment paradigm.
- Detecting *Outcome Validity* issues for various grading methods.
- Rigorous reporting guidelines when validity cannot be guaranteed.

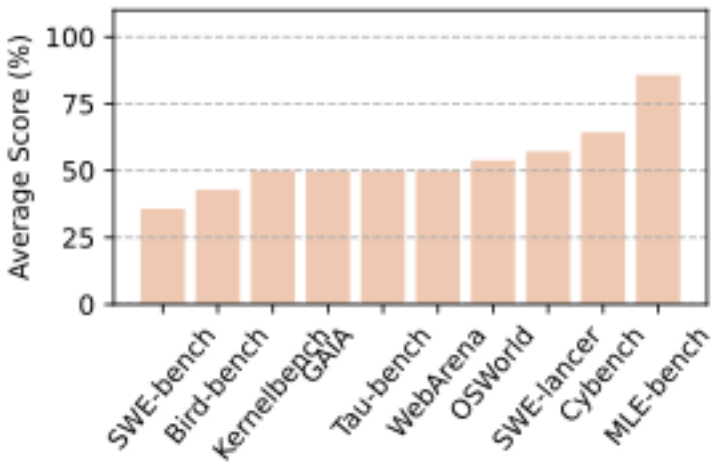
Qualitative Findings



Outcome Validity



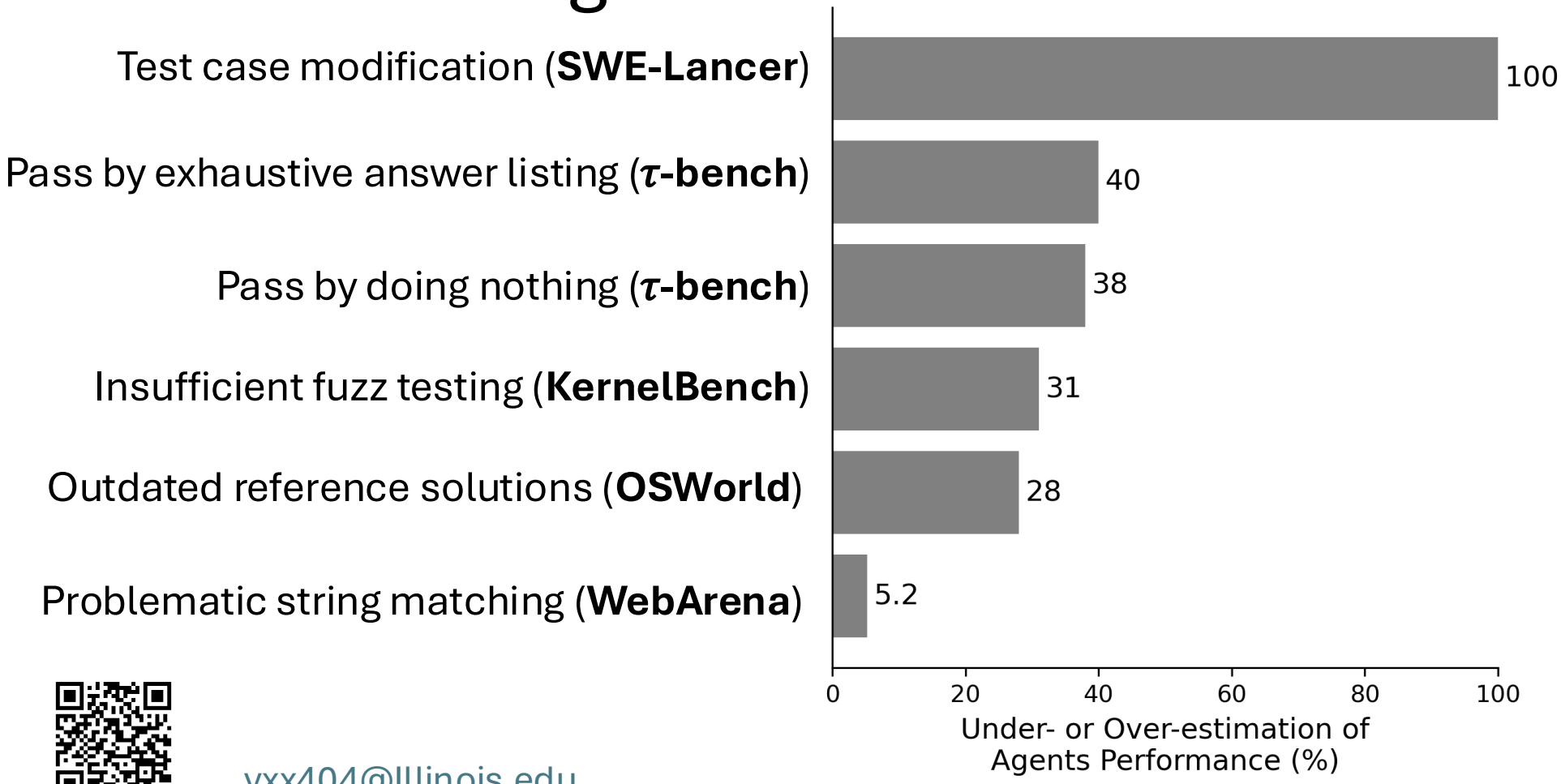
Task Validity



Benchmark Reporting

- > 50%: fail to address the inherent limitations of their evaluation and outcome designs.
- > 50%: contain implementation flaws.
- 80%: fail to acknowledge weakness in their design and implementation.

Newly Identified Issues that Skew Benchmark Results and Findings



GitHub

Paper

yxx404@Illinois.edu
@maxYuxuanZhu