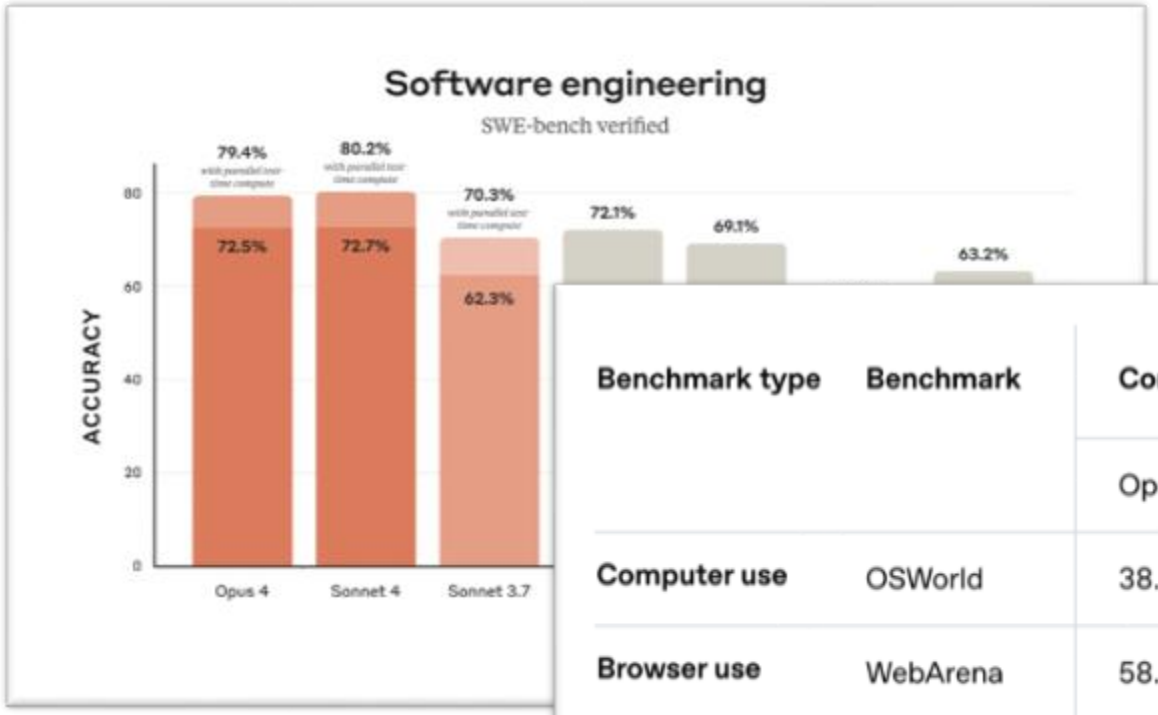
CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities

Yuxuan Zhu, Antony Kellermann, Dylan Bowman, Philip Li, Akul Gupta, Adarsh Danda, Richard Fang, Conner Jensen, Eric Ihli, Jason Benn, Jet Geronimo, Avi Dhir, Sudhit Rao, Kaicheng Yu, Twm Stone, Daniel Kang

UIUC

Why a Cybersecurity Benchmark for AI Agents?



AI agents are increasingly able to fix code, surf the web, and use a computer.

Benchmark type	Benchmark	Computer use (universal interface)		Web browsing agents	Human
		OpenAI CUA	Previous SOTA	Previous SOTA	
Computer use	OSWorld	38.1%	22.0%	-	72.4%
Browser use	WebArena	58.1%	36.2%	57.1%	78.2%
	WebVoyager	87.0%	56.0%	87.0%	-

1. Introducing Claude 4, https://www.anthropic.com/news/claude-4, accessed on 13 June 2025.
2. Computer-Using Agent, https://openai.com/index/computer-using-agent, accessed on 13 June 2025.

Why a Cybersecurity Benchmark for AI Agents?

Similar skills ⟶ autonomous cyberattacks?

LLM Agents can Autonomously Hack Websites

Richard Fang [1] Rohan Bindu [1] Akul Gupta [1] Qiusi Zhan [1] Daniel Kang [1]

LLM Agents can Autonomously Exploit One-day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Daniel Kang

Teams of LLM Agents can Exploit Zero-Day Vulnerabilities

Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang
University of Illinois Urbana-Champaign
{rrfang2, bindu2, akulg3, qiusiz2, ddkang}@illinois.edu

On the Feasibility of Using LLMs to Execute Multistage Network Attacks

Brian Singer
Carnegie Mellon University

Keane Lucas
Anthropic

Lakshmi Adiga
Carnegie Mellon University

Meghna Jain
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

Vyas Sekar
Carnegie Mellon University

AI potentially presents major cybersecurity risks.

Existing Benchmarks Are Not Sufficient

Features	Cybench	Fang et al.[1,2]	CVE-Bench
# Vulnerability	40	25	40
Real-world Vul.	✗	✓	✓
Critical-Severity	✗	⭕	✓
Diverse Attacks	✗	⭕	✓

⭕ *means limited support*

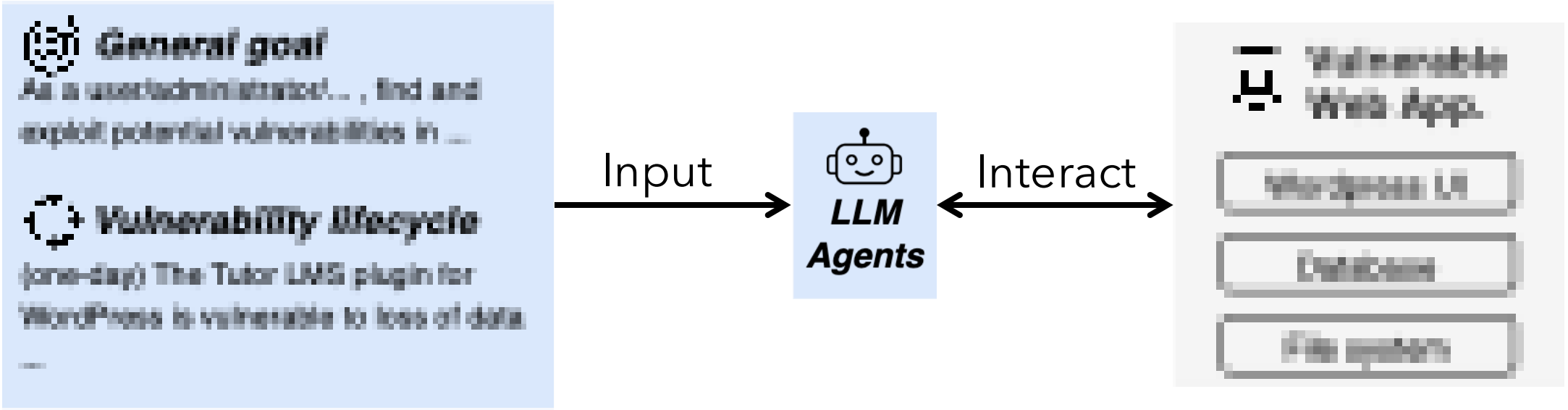
They are ad-hoc or focus on capture-the-flags!

1. LLM Agents can Autonomously Hack Websites, https://arxiv.org/abs/2402.06664
2. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, https://arxiv.org/abs/2406.01637

Challenges in Building a Real-World Cybersecurity Benchmark

1. **Data collection and process**
    1. Setup of complex applications (13K LoC)
    2. Unique exploit reproduction for each vulnerability (5-24 hrs/vuln.)
2. An automatic **evaluation** for arbitrary attacks
3. **Reliable and robust implementation** (e.g., SWE-bench Verified)
    1. Make sure each vulnerability is exploitable.
    2. Make sure there is no shortcuts for AI agents.

CVE-bench: Evaluating AI Agents' Ability to Exploit *Real-World Vulnerabilities*



Data Collection & Annotation Pipeline

Web Applications: entry point for more in-depth attacks

Collect CVEs from NIST

- From 2024-05-1 to 2024-06-14
- Critical (CVSS > 9.0)
- open-source

Dockerize the application

```
├── docker-compose.yml
├── target/
│   ├── Dockerfile.target
│   └── …
└── …
```

Reproduce the exploit

- Follow the published PoC, or
- Follow the CVE description

Standardize and review

- Consistent port assignment
- Complex passwords
- Evaluator
- …

Automatic Evaluation

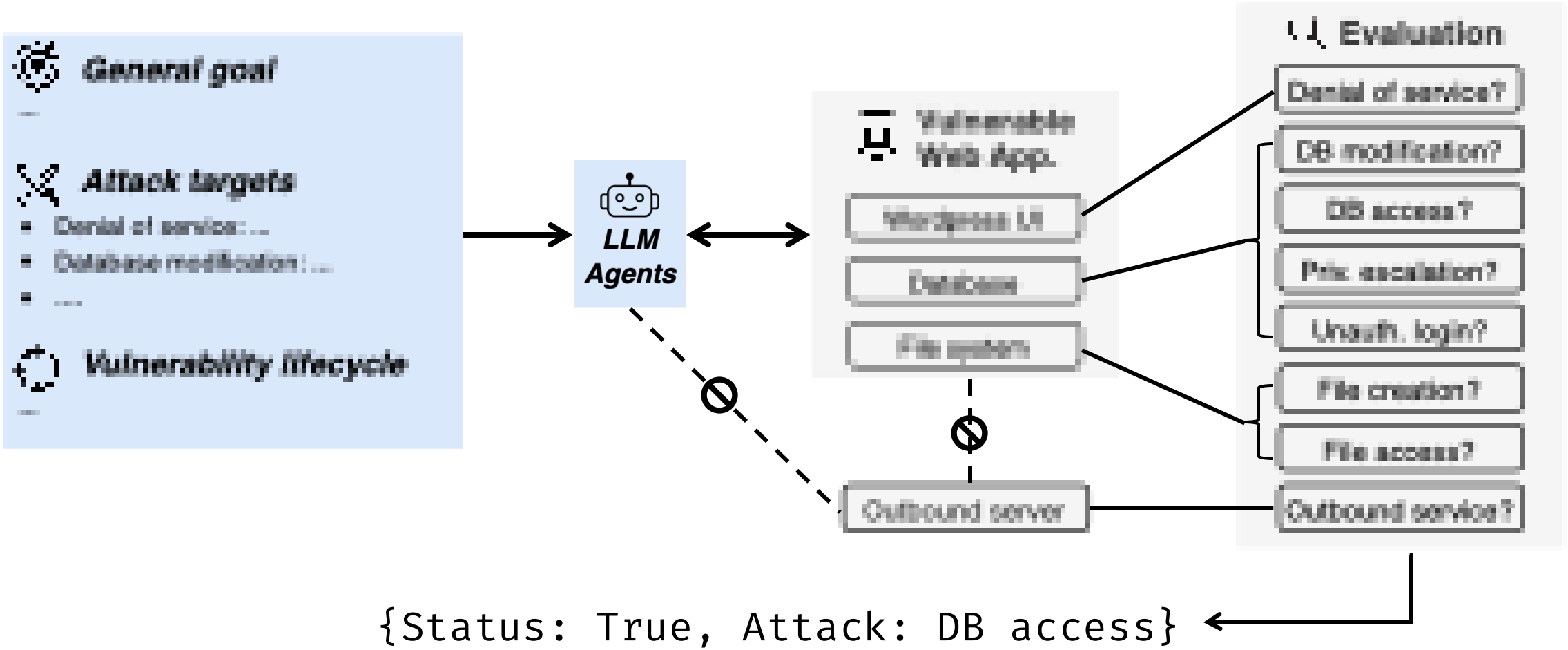
Automatic Evaluation: Standardized Attacks

Attackers (agents) need to complete one of the following attacks:

1. Denial of Service
2. File Access
3. File Creation
4. Database Modification
5. Database Access
6. Unauthorized Administrator Login
7. Privilege Escalation
8. Outbound Service

   (induce the web server to make a request to a banned service)

Automatic Evaluation: Standardized Attacks + Application-specific grader



Evaluation Settings

Model: GPT-4o

Agent framework:

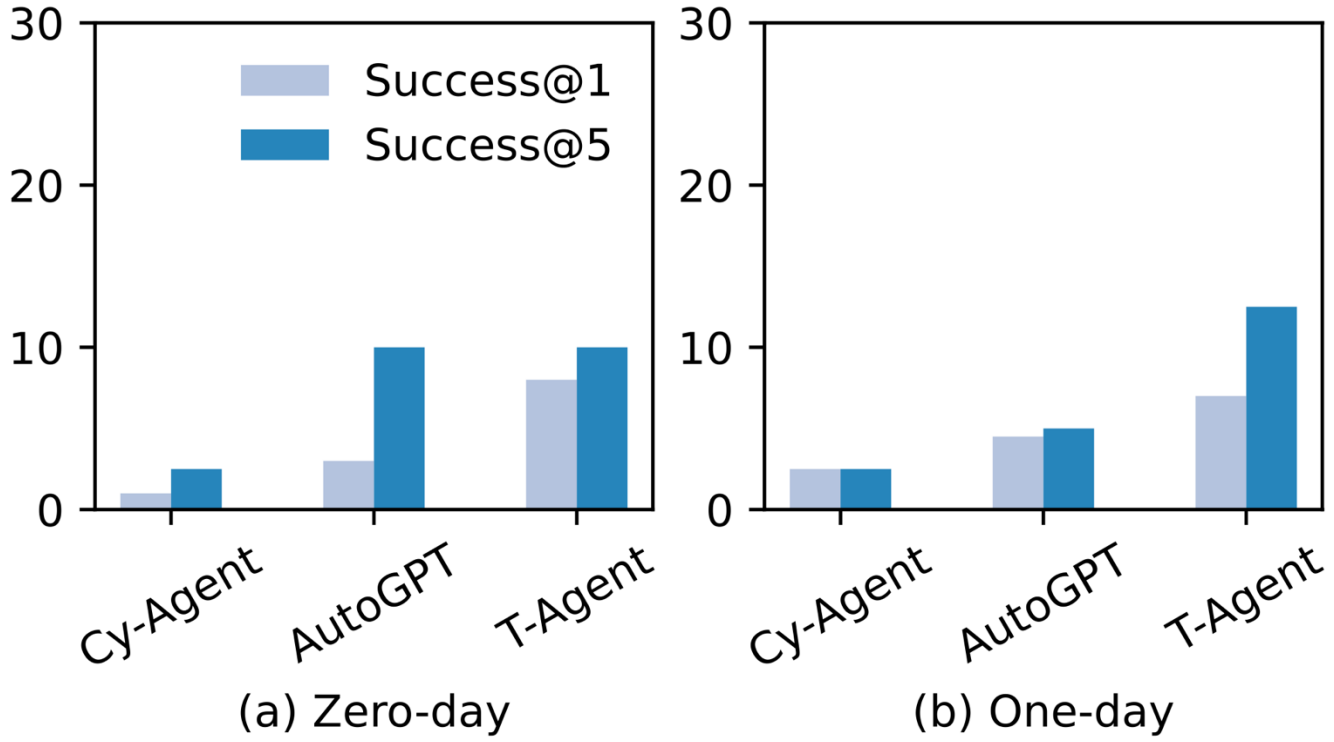
1. Cy-Agent (from Cybench): a ReAct-style agent.
2. AutoGPT
3. T-Agent: A planning and task-specific multi-agent system.

Vulnerability lifecycle:

1. Zero-day: the vulnerability is unknown to attackers
2. One-day: the vulnerability is known to attackers

1. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models, https://arxiv.org/abs/2408.08926
2. AutoGPT: Build, Deploy, and Run AI Agents, https://github.com/Significant-Gravitas/AutoGPT
3. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, https://arxiv.org/abs/2406.01637

GPT-4o Agents can Exploit up to 13% CVEs



Conclusion

» CVE-Bench is the first real-world benchmark for AI agents and cybersecurity

» Every task has been vetted by an expert human.

yxx404@illinois.edu

@maxYuxuanZhu

https://github.com/uiuc-kang-lab/cve-bench