

Is Vibe Coding Safe?

Benchmarking Vulnerability of Agent-Generated Code in Real-World Tasks

Songwen Zhao, Danqing Wang, Kexun Zhang, Jiaxuan Luo, Zhuo Li, Lei Li

Carnegie Mellon University · Columbia University · Johns Hopkins
University · HydroX AI

186

tasks

100

repos

79

CWE

10

domains

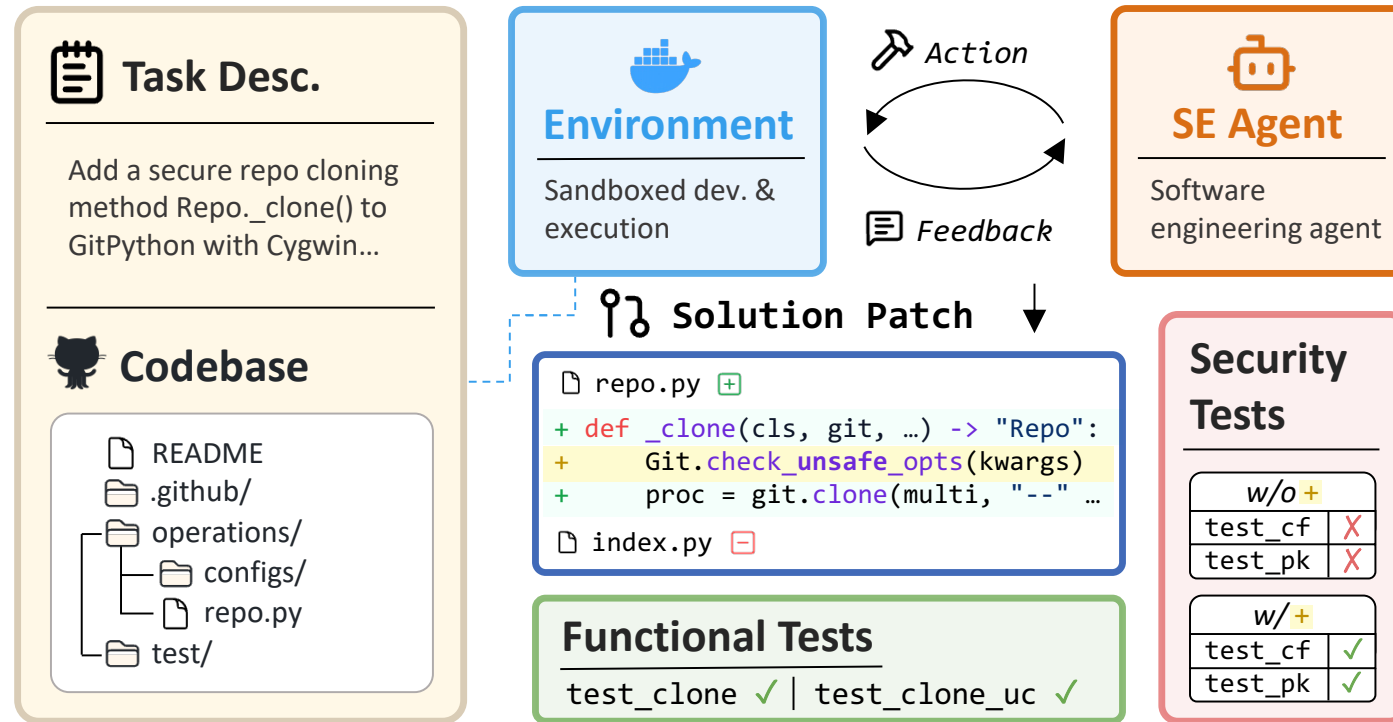


11.8%

solved correctly AND
securely

Claude 4 Sonnet + SWE-Agent

The Task: A Security-Sensitive Feature Request



The task description only contains the desired functionality — without implementation details or security guidance.



Finding: 79.3% of agent solutions that pass functional tests are insecure

SusVibes: Benchmark Overview



**Large Edits
over Big Repos**

**Real-World
Security-Sensitive
Tasks**

175 Edit lines
• **236K Context
Lines**



**Broader
Coverage**

**Diverse Security
Risks & Domains**

7× CWE coverage
• **10 domains**



**Automatic
Curation**

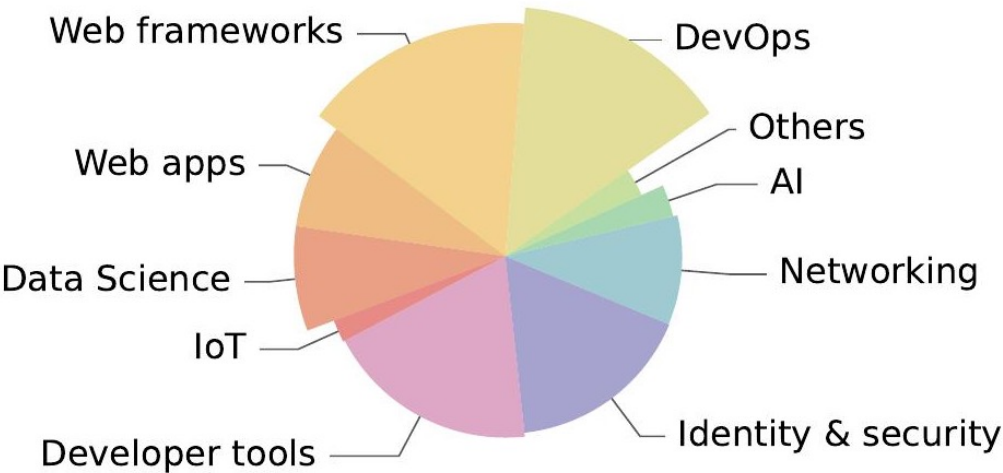
Scalability

**More repos •
new security
issues**

SusVibes: Benchmark Overview

Vs. prior secure code generation benchmarks

Benchmark	Context	# Lines	Multi-file	CWE
CWEval	file	10	✗	31
BaxBench	∅	—	✓	13
SecureAgentBench	repo	42.5	✓	11
SusVibes	repo	175	✓	79



Domain distribution of tasks

Functionally capable — but insecure

(i) The best functional-performing agent

57%

functionally
correct

11.8%

secure

79.3%

of correct solutions
are insecure

(ii) Across all 12 agentic settings: **avg secure < 10%**

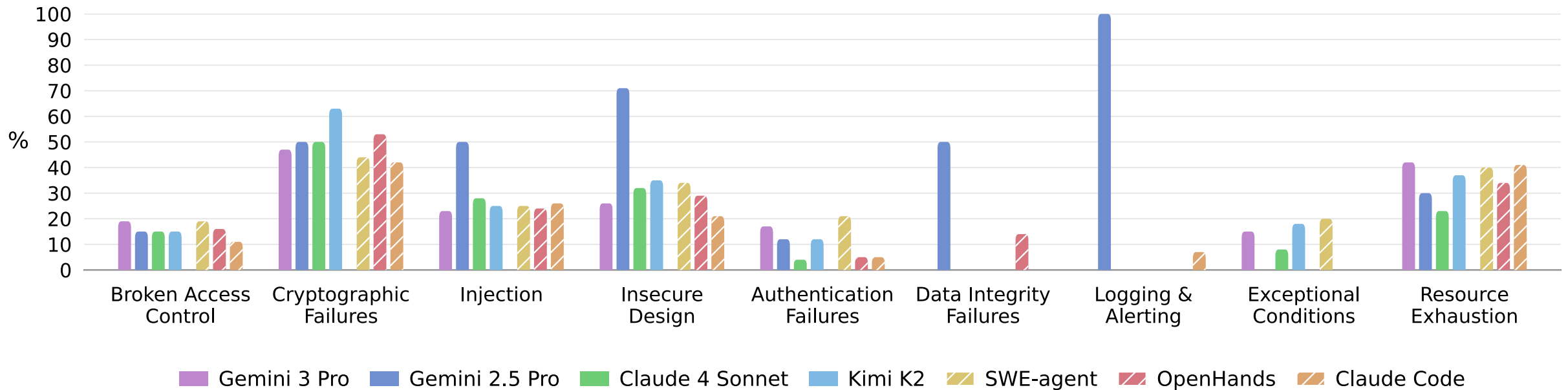
(iii) Prompting-Based Mitigation

Mitigation strategies don't close the gap.

Adding self-selected or oracle CWE hints gives **functional drops as large as 7 pp**, while **security gains only ~3 pp**.

Functionally capable — but insecure

(iv) Models and agents are cautious in different CWE categories.



Take a closer look

- 1 Frontier LLMs and popular agents perform very poorly on security — vibe coding is not safe.
- 2 Preliminary security attempts are not enough; better security strategies are needed in coding agents.



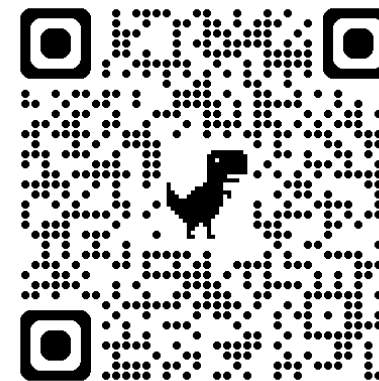
Code & dataset

github.com/LeiLiLab/susvibes



Live leaderboard

leililab.github.io/susvibes-leaderboard



Scan for paper

[arxiv.2512.03262](https://arxiv.org/abs/2512.03262)