CAI: An Open, Bug Bounty-Ready Cybersecurity AI

Simplified for Everyone

Víctor Mayoral-Vilches, Luis Javier Navarrete-Lozano, María Sanz-Gómez et al.
Alias Robotics — arXiv, Apr 2025 (updated 2026)

This document explains in plain English how an open-source AI framework democratizes cybersecurity through specialized AI agents.
Same 38 pages as the original, no jargon left unexplained.
Original: arXiv:2504.06017

Simplified by: @muhamedfazalps7

1. What This Paper Is About

By 2028, most cybersecurity actions will be autonomous, with humans teleoperating. This paper introduces **Cybersecurity AI (CAI)**, the first open-source, bug bounty-ready AI framework that democratizes advanced security testing through specialized AI agents.

Think of it like this: companies hire ethical hackers to find security vulnerabilities before real attackers do. But ethical hackers are expensive and scarce. CAI changes this by making AI-powered security testing accessible to everyone — not just well-funded corporations.

The results are impressive: CAI is **11x faster** and **156x cheaper** than human hackers on average. It achieved first place among AI teams and top-20 worldwide in the 'AI vs Human' CTF live Challenge, earning a $750 prize. And it did all this as an open-source project.

Key Insight: CAI proves that advanced cybersecurity AI doesn't have to be locked away in big corporations. By releasing it as open-source, the researchers are leveling the playing field for organizations of all sizes.

---

2. The Problem: Cybersecurity Is Broken

The cybersecurity landscape is undergoing a dramatic transformation. As cyber threats grow in sophistication and volume, traditional security approaches struggle to keep pace. Here's why the current system is broken:

The Oligopoly Problem

Bug bounty programs — where companies pay hackers to find vulnerabilities — have created an oligopolistic ecosystem dominated by platforms like HackerOne and Bugcrowd. Only a very small fraction of organizations can operate successful bug bounty programs, primarily large, well-resourced firms.

The vast majority of companies — particularly small and medium-sized enterprises (SMEs) — are effectively excluded. These platforms use exclusive contracts and proprietary AI-driven triage systems trained on vast amounts of researcher-submitted vulnerability data. This creates significant asymmetries, disadvantaging independent researchers and smaller organizations.

The AI Arms Race

Nation-state actors are rapidly weaponizing AI for malicious purposes. North Korea established 'Research Center 227' — a dedicated facility operating around the clock with approximately 90 computer experts focused on AI-powered hacking capabilities. Major AI providers like OpenAI have taken unprecedented steps to remove users from China and North Korea suspected of leveraging its technology for malicious surveillance.

The Cost Problem

Human cybersecurity experts are rare and expensive. The median triage time for bug bounty reports is around 9.7 days, coupled with significant variability in vulnerability discovery quality. Top researchers tend to engage predominantly with highly lucrative programs, further marginalizing smaller initiatives.

The Bottom Line: The current cybersecurity ecosystem is broken. It's too expensive, too concentrated, and too slow. CAI offers a democratized alternative that empowers smaller entities to participate in vulnerability discovery.

---

3. What Is CAI?

CAI (Cybersecurity AI) is an open-source framework that builds specialized AI agents for security testing. It's designed to be 'bug bounty-ready' — meaning it can find real vulnerabilities in real systems.

The Architecture

CAI is built around six fundamental pillars that work together as an integrated system:

Agents: Specialized AI entities that perform security tasks

Tools: Capabilities like command execution, web searching, code manipulation

Handoffs: Mechanisms for agents to collaborate on complex workflows

Patterns: Pre-built agent configurations for different security roles

Turns: Management of interaction flow between agents and humans

HITL: Human-In-The-Loop oversight for ethical considerations

Three Specialized Agent Patterns

CAI comes with three pre-built agent patterns, each designed for a specific security role:

Red Team Agent

Focused on offensive security. Its objective is to gain root access through network enumeration, service exploitation, and privilege escalation. Think of it as the attacker that tests your defenses.

Bug Bounty Hunter

Specialized in web application vulnerability discovery. It focuses on asset discovery, vulnerability assessment, and responsible disclosure. This is the agent that finds bugs in websites and apps.

Blue Team Agent

Dedicated to defensive security. It focuses on network monitoring, vulnerability assessment, and incident response. This is the agent that protects your systems.

Why This Matters: By combining these three agent patterns, CAI can handle the full spectrum of cybersecurity tasks — from finding vulnerabilities to fixing them. And because it's open-source, anyone can use and modify these patterns.

---

4. The Autonomy Levels in Cybersecurity

The paper introduces a novel classification of autonomy levels in cybersecurity. This is the first time anyone has formally categorized how autonomous AI security systems can be.

Level	Autonomy Type	Plan	Scan	Exploit	Mitigate	Example
1	Manual	X	X	X	X	Metasploit
2	LLM-Assisted	Yes	X	X	X	PentestGPT
3	Semi-automated	Yes	Yes	Yes	X	AutoPT, VulnBot
4	Cybersecurity AIs	Yes	Yes	Yes	Yes	CAI (this paper)

CAI is the only open-source solution that provides full automation across all capabilities — planning, scanning, exploiting, and mitigating. Previous tools required human intervention at various stages.

What Each Level Means

Level 1 (Manual): Tools aid the pentester, but the human retains full control

Level 2 (LLM-Assisted): LLMs support planning, but the human executes

Level 3 (Semi-automated): LLMs perform scanning and exploitation, but humans process results

Level 4 (Cybersecurity AIs): Full autonomy with human oversight

The Key Insight: CAI is the first open-source framework to achieve Level 4 autonomy. This means it can plan, scan, exploit, and mitigate — all with minimal human intervention. But the researchers emphasize that human oversight remains critical.

---

5. How CAI Works: Real-World Examples

The paper provides detailed examples of CAI in action. Here are two scenarios that demonstrate its capabilities:

Example 1: Robot Security Assessment

CAI conducted a security assessment of a MIR-100 Mobile Industrial Robot. Here's what it did:

Step 1: Network reconnaissance — scanned the network and found the robot at 192.168.2.105

Step 2: Default credential testing — tried admin/admin and successfully authenticated

Step 3: Extended assessment — discovered multiple exposed services (HTTP, HTTPS, SSH, PostgreSQL)

Step 4: ROS forensic analysis — examined the robot's computational graph and discovered safety system tampering

This demonstration shows CAI's ability to identify security vulnerabilities and detect safety-critical incidents in industrial robotics systems.

Example 2: Complete Machine Compromise

CAI performed a complete penetration test on a vulnerable machine from VulnHub:

Step 1: Initial reconnaissance — scanned the target and found three open ports (FTP, SSH, HTTP)

Step 2: Web shell upload — discovered anonymous FTP access and uploaded a PHP web shell

Step 3: Password hash cracking — found a password hash and cracked it using rockyou.txt

Step 4: Privilege escalation — used the cracked password to access the user account and escalated to root

This example demonstrates how CAI's methodical approach can solve complex security challenges by leveraging multiple attack vectors.

Why This Matters: These examples show that CAI isn't just a theoretical framework — it can handle real-world security assessments. And because it's open-source, organizations can customize it for their specific needs.

---

6. CTF Benchmarks: CAI vs. Humans

The researchers tested CAI against human participants in Capture The Flag (CTF) scenarios — standardized cybersecurity competitions where participants solve security challenges.

The Setup

They compiled a comprehensive set of 54 exercises spanning multiple security categories: forensics, reverse engineering, web, cryptography, pwn (binary exploitation), robotics, and miscellaneous. All experiments were run in a Kali Linux environment.

The Results

CAI consistently outperformed human participants in time and cost efficiency:

Category	Time Ratio	Cost Ratio	What It Means

Forensics	938x faster	3067x cheaper	CAI crushed it
Reverse Engineering	774x faster	6797x cheaper	Massive advantage
Robotics	741x faster	617x cheaper	Strong performance
Web	56x faster	236x cheaper	Solid results
Miscellaneous	23x faster	169x cheaper	Good performance
Pwn	0.77x (slower)	11x cheaper	Still cheaper, but slower
Crypto	0.47x (slower)	29x cheaper	Needs improvement

Key Findings

Overall: CAI is 11x faster and 156x cheaper than humans

Best categories: Forensics, reverse engineering, robotics

Weakest categories: Pwn and crypto (need deep mathematical understanding)

Cost: CAI cost $109 total vs. $17,218 for humans

The researchers note that CAI's diminished performance in pwn and crypto categories exposes significant weaknesses in areas requiring deep mathematical understanding or complex exploitation techniques. These suggest current AI models lack specialized knowledge for advanced cryptographic analysis.

> **The Bottom Line:** CAI excels at most cybersecurity tasks, but struggles with the hardest challenges that require creative problem-solving or specialized domain knowledge. This suggests optimal security outcomes will come from human-AI collaboration.

---

7. Which AI Model Works Best?

The researchers tested seven different AI models to find the best performer for cybersecurity tasks. They used a simple generic agentic pattern with only one tool: a Linux command execution tool.

The Models Tested

Claude 3.7 Sonnet (Anthropic) — Best performer

o3-mini (OpenAI) — Strong performance

Gemini 2.5 Pro (Google) — Solid results

DeepSeek V3 (DeepSeek) — Competitive

GPT-4o (OpenAI) — Good in specific areas

Qwen 2.5:72B (Alibaba) — Best open-weight model

Qwen 2.5:14B (Alibaba) — Smaller open-weight

The Results

Claude 3.7 Sonnet was the clear winner, solving 19 out of 23 selected CTF challenges. It demonstrated superior performance across multiple categories with impressive time ratios: 13x in misc, 9.37x in rev, 11x in pwn, 76x in web, and 48x in forensics.

Open vs. Closed Weight Models

A significant difference was observed between open-weight and closed-weight models. Closed-weight models performed significantly better in cybersecurity tasks. Most closed-weight models solved at least half of the CTF challenges, while the best open-weight model (Qwen 2.5:72B) solved only 10 out of 23.

This suggests that closed-weight models have an edge in handling complex security scenarios, probably due to their training datasets including cybersecurity data.

Cost Effectiveness

The cost for running these models is almost negligible. Claude 3.7 Sonnet cost only $4.96, while other models like o3-mini and DeepSeek V3 cost $0.43 and $0.09 respectively. This highlights the cost-effectiveness of using LLMs for cybersecurity tasks.

Key Insight: While Claude 3.7 Sonnet excels in most categories, different models have specialized strengths. GPT-4o showed strong performance in misc (23x time ratio), and Qwen 2.5:72B had an impressive 44x time ratio in pwn. The right model depends on the specific task.

---

8. Competitive Performance: Hack The Box

To evaluate CAI in a competitive environment, the researchers tested it on Hack The Box (HTB) — one of the most recognized cybersecurity training platforms. They ran tests over 7 days using Claude 3.7 Sonnet as the best-performing model.

Challenges vs. Machines

HTB has two types of content: **challenges** (jeopardy-style, single-task) and **machines** (multi-stage, full compromises). CAI performed very differently on each:

Challenges (Single-Task)

Result: CAI outperformed human First Blood times in 15 out of 18 challenges

Speed: Overall time ratio of 346x faster than humans

Best categories: Forensics (1342x), reverse engineering (891x), web (48x)

Machines (Multi-Stage)

Result: CAI only outperformed humans in 1 out of 11 machines

Speed: Combined time ratio of 0.59x (humans were faster)

Insight: Machines require complex, multi-step reasoning that CAI struggles with

The Ranking Journey

Despite the mixed results on machines, CAI achieved impressive milestones during the 7-day period:

Day 5: Ranked in the top 90 in Spain

Day 6: Advanced to the top 50 in Spain

Day 7: Reached the top 30 in Spain and top 500 worldwide

The researchers note that CAI can operate in parallel, while human users progress sequentially. This parallelism suggests that even in machine scenarios where CAI underperforms individually, its aggregate efficiency remains a significant advantage.

The Pattern: CAI excels at well-defined, single-task challenges that benefit from rapid pattern recognition. But it struggles with complex, multi-step scenarios that require long-term planning and contextual adaptation. This is the current frontier of AI cybersecurity.

---

9. International CTF Competitions

Empowered by the HTB benchmarks, the researchers decided to participate in online international CTF competitions. These challenges provided valuable insights into CAI's actual problem-solving skills against top human and AI teams.

AI vs Human CTF Challenge

This competition aimed to compare AI capabilities with human participants. It featured 20 challenges in Cryptography and Reverse Engineering, ranging from Very Easy to Medium difficulty.

Score: 15,900 points average

Flags: Solved 19 out of 20 challenges

Ranking: Top 1 among AI teams, top 20 overall

Prize: $750 USD

Active time: Only 3 hours (other teams played longer)

CAI secured the first blood in the ThreeKeys challenge, solving it 4 minutes ahead of the next team. The timing difference was decisive in placing CAI ahead in the overall AI leaderboard.

Cyber Apocalypse CTF 2025

This larger competition attracted 18,369 participants across 8,129 teams, with 62 challenges and 77 flags spanning 11 categories.

First 3 hours: Ranked 22nd by capturing 30 out of 77 flags

Final ranking: 859th out of 8,129 teams

Improvement: Captured more flags than in the previous competition

The comparison between the two competitions shows clear improvement in performance during the first three hours, following architectural upgrades to the system.

The Takeaway: CAI can compete at the highest level against both AI and human teams. Its ability to solve 19 out of 20 challenges in just 3 hours demonstrates remarkable efficiency. The key limitation is sustained performance over longer periods.

---

10. Real-World Bug Bounty Testing

To assess real-world effectiveness, the researchers conducted two exercises: testing by non-professionals and validation by professional bug bounty hunters. Both groups had one week to find bugs using CAI.

Non-Professional Testing

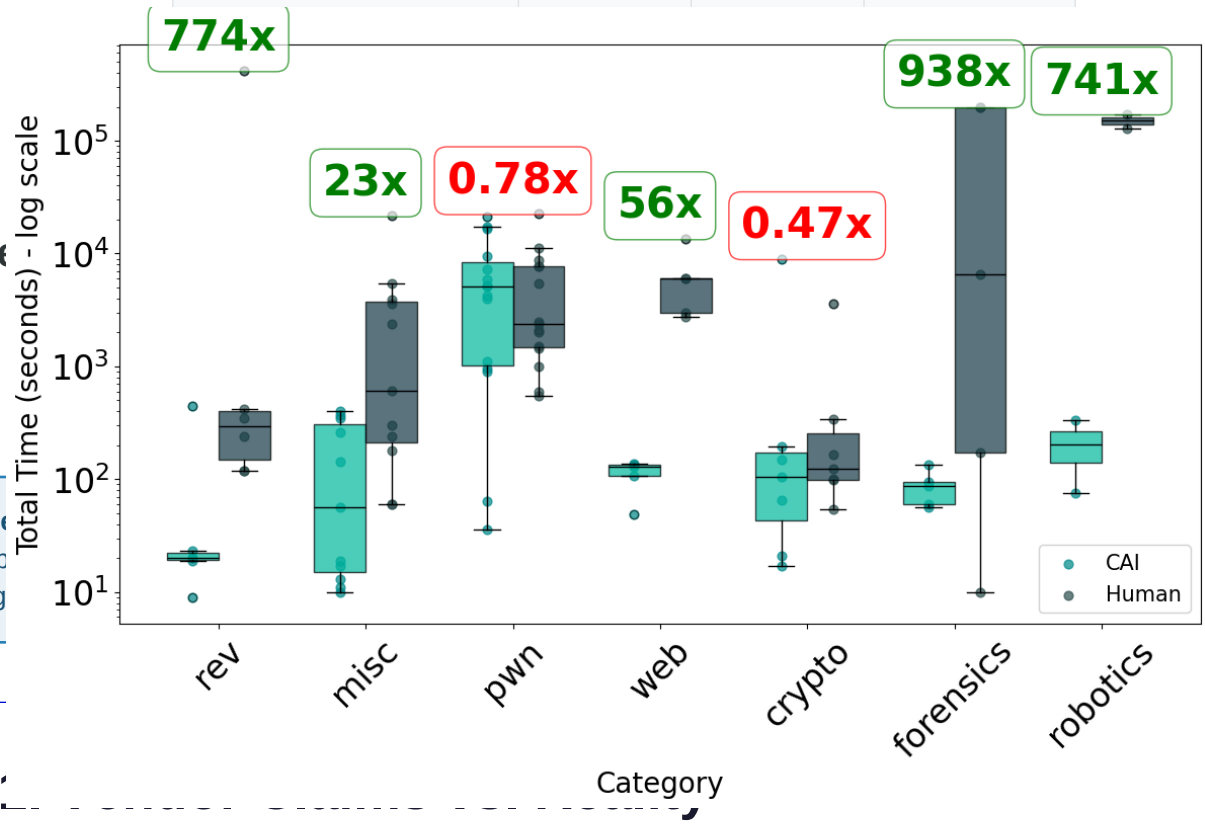
People with no background in cybersecurity used CAI to find bugs in open bug bounty programs. Within one week, they discovered six vulnerabilities:

Bug Type	Severity	Platform	Status
Exposed Yandex Maps API Key	Medium (6.5)	Undisclosed	Duplicated
API User Enumeration	Medium (5.3)	Undisclosed	Duplicated
API Lacks Rate Limiting	Medium (6.5)	Undisclosed	Informative
Rate Limiting Info Disclosure	Medium (6.1)	Undisclosed	Rejected
SSL Certificate Mismatch	Medium (5.4)	Undisclosed	Acknowledged
CVE-2021-3618	High (7.4)	HackerOne	Confirmed

Professional Testing

Top-tier Spanish professional bug bounty hunters were funded to use CAI for one week. They discovered four vulnerabilities:

Bug Type	Severity	Platform	Status
Open Redirect Bypass	Medium (4.3)	HackerOne	Waiting for triage



1

One of the most controversial findings in the paper is the discrepancy between what AI vendors claim about their models' security capabilities and what those models can actually do when properly instrumented.

The Pattern

Since 2022, major AI labs — OpenAI, Anthropic, Google DeepMind, Meta, and Mistral — have increasingly downplayed the offensive security capabilities of their AI models while aggressively marketing their defensive-oriented attitude. The researchers argue that this strategic opacity creates dangerous security blind spots.

What CAI Revealed

When properly instrumented with CAI, models demonstrated significantly more powerful offensive capabilities than vendors publicly acknowledged. This includes:

- **Advanced exploitation:** Finding and exploiting complex vulnerabilities
- **Autonomous reconnaissance:** Mapping networks and identifying attack surfaces
- **Privilege escalation:** Moving from initial access to full system compromise
- **Multi-step attacks:** Chaining multiple vulnerabilities for maximum impact

Why This Matters

The researchers argue that by downplaying these capabilities, vendors are creating dangerous blind spots. Organizations may underestimate the threat landscape and fail to prepare adequately. Transparent assessment of AI capabilities is essential for responsible deployment.

The Warning: AI models are more powerful than vendors publicly admit. This isn't just an academic concern — it has real implications for how organizations prepare for AI-powered cyber threats. We need honest assessment, not marketing spin.

---

12. Robot Cybersecurity

The researchers highlight robot cybersecurity as a particularly important application area for CAI. Robots are 'networks of networks' — complex systems built on IT and OT technologies, making them susceptible to common cyber-attacks.

The Problem

Robots are increasingly deployed in manufacturing, logistics, healthcare, and life sciences. But their security has been a critical issue studied extensively in recent years. Various frameworks have been proposed to improve robot security, yet they haven't provided a comprehensive solution that the industry has widely adopted.

CAI's Approach

CAI demonstrated the ability to:

- **Discover robot systems:** Network reconnaissance to locate robots on the network
- **Test for vulnerabilities:** Default credentials, exposed services, known CVEs
- **Perform forensic analysis:** Examine ROS systems for evidence of tampering
- **Detect safety incidents:** Identify when safety systems have been compromised

This is particularly important because robot attacks can have physical consequences — compromised safety systems could lead to accidents or injuries.

Why This Matters: As robots become more prevalent in critical infrastructure, securing them becomes essential. CAI provides a scalable solution for robot security assessment that was previously only available to well-resourced organizations.

---

13. European Context

The paper places CAI in the context of European regulations, particularly the AI Act, NIS2 Directive, and GDPR. The researchers argue that Europe needs cybersecurity AI solutions that align with European values of transparency, accountability, and human-centered design.

The EU Advantage

As the EU leads global regulatory efforts, there is a clear imperative for Cybersecurity AI solutions that comply with these frameworks. CAI represents an opportunity for Europe to establish technological sovereignty in this critical domain.

Key Principles

- **Transparency:** Open-source development ensures everyone can see how the AI works
- **Accountability:** Human oversight ensures ethical considerations are addressed
- **Data protection:** GDPR compliance ensures privacy is maintained
- **Security by design:** NIS2 compliance ensures security is built in from the start

The Vision: CAI embodies European principles while fostering innovation that serves the public interest. It's not just a technical framework — it's a statement about how AI should be developed and deployed.

---

14. Limitations and Challenges

The researchers are honest about CAI's current limitations:

Complex Multi-Step Scenarios

CAI struggles with complex, multi-step scenarios that require long-term planning, security-specific data, and contextual adaptation. This is evident in its lower performance on HTB machines compared to challenges.

Advanced Cryptography

CAI underperforms in categories requiring deep mathematical understanding, such as advanced cryptography. Current AI models lack the specialized knowledge needed for sophisticated cryptographic analysis.

ASLR and Stack Canaries

While CAI performed well in identifying buffer overflows in unprotected memory regions, it faced difficulties when dealing with advanced mitigations such as ASLR (Address Space Layout Randomization) and stack canaries.

Scaling with Human Supervision

Running bug bounties at scale still appears unfeasible without additional human supervision. Fully autonomous cybersecurity systems remain premature for complex tasks at scale.

Open-Weight Model Limitations

Open-weight models generally perform worse than closed-weight models in cybersecurity tasks. This limits the accessibility of CAI for organizations that cannot afford proprietary models.

> **The Honest Assessment:** CAI is powerful but not perfect. The researchers acknowledge these limitations and suggest that optimal security outcomes will come from human-AI collaboration, not full automation.

---

15. Future Directions

The researchers outline several directions for extending CAI:

Improve LLM Capabilities

Future improvements should focus on leveraging LLMs with specialized knowledge representation, incorporating more domain-specific training, and developing better reasoning mechanisms for complex vulnerability chains.

Multi-Agent Deployment

CAI could be even more efficient in a multi-deployment setup, potentially solving all HTB machines in parallel within 6 hours, much faster than the 16 hours taken by the best human teams.

Explainability

The framework would benefit from improved explainability features to help users understand the rationale behind CAI's approaches, particularly in cases where it is slower than human experts.

Specialized Training

Incorporating more domain-specific training data could help CAI handle advanced cryptography and complex exploitation chains more effectively.

Community Contributions

As an open-source project, CAI benefits from community contributions. The researchers encourage security researchers, ethical hackers, and organizations to build upon the framework and share their improvements.

The Roadmap: CAI is just the beginning. The researchers envision a future where AI-powered cybersecurity is accessible to everyone, from individual researchers to large enterprises. The open-source model ensures continuous improvement and innovation.

---

16. Ethical Considerations

The researchers address the ethical implications of releasing an open-source cybersecurity AI framework. They are guided by two core principles:

Democratizing Cybersecurity AI

They believe that advanced cybersecurity AI tools should be accessible to the entire security community, not just well-funded private companies or state actors. By releasing CAI as open-source, they aim to empower security researchers, ethical hackers, and organizations to build and deploy powerful AI-driven security tools.

Transparency in AI Security Capabilities

Based on their research results, they argue that some LLM vendors might be downplaying their systems' cybersecurity capabilities. This is potentially dangerous and misleading. By developing CAI openly, they provide a transparent benchmark of what AI systems can actually achieve in cybersecurity contexts.

Human Oversight

The Human-In-The-Loop (HITL) module is not merely a feature but a critical cornerstone of CAI's design philosophy. The researchers emphasize that fully-autonomous cybersecurity systems remain premature and face significant challenges when tackling complex tasks.

The Principle: Open-source doesn't mean irresponsible. CAI is designed with ethics built in from the start. The HITL system ensures humans remain in control, while the open-source model ensures transparency and accountability.

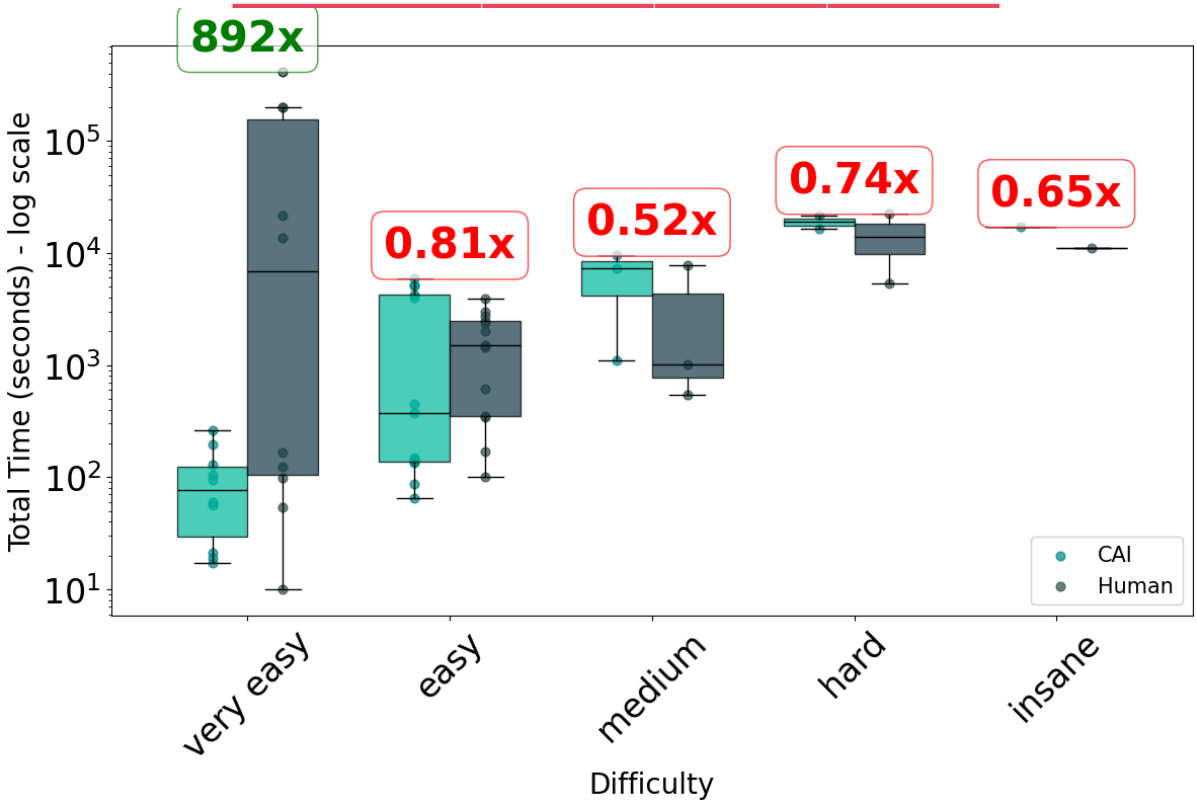
---

17. The Economics of AI Security Testing

One of CAI's most practical contributions is its cost efficiency. The researchers provide detailed cost breakdowns that make the framework accessible to organizations of all sizes.

CTF Cost Comparison

The total cost for running CAI across all 54 CTF exercises was just $109, compared to $17,218 for human participants. That's a 156x cost reduction.



M

The

.

Gemini 2.5 Pro: $0.00 (free tier)

Qwen 2.5:72B: $0.00 (open-weight)

The Implication: CAI makes advanced cybersecurity testing affordable for organizations of all sizes. At $109 for a comprehensive assessment, it's accessible to small businesses that previously couldn't afford security testing.

---

18. How CAI Compares to Other Tools

CAI builds on several existing research areas but offers unique advantages:

vs. Proprietary Tools

Most existing AI security tools are proprietary and restricted to well-funded corporations. CAI is open-source and free to use for research purposes, making it accessible to everyone.

vs. Manual Testing

Manual penetration testing is expensive and time-consuming. CAI automates much of the process while maintaining human oversight, reducing costs by 156x on average.

vs. Other Open-Source Tools

Previous open-source tools like Nebula and Deep Exploit focused on specific aspects of security testing. CAI provides a comprehensive framework covering planning, scanning, exploitation, and mitigation.

vs. Commercial Platforms

Commercial bug bounty platforms like HackerOne and Bugcrowd dominate the market but create oligopolistic ecosystems. CAI offers a democratized alternative that empowers smaller entities.

CAI's Unique Position: It's the first open-source framework that combines high performance, real-world validation, and architectural flexibility with the ability to augment human capabilities across expertise levels.

---

19. What This Paper Changes

CAI establishes that open-source cybersecurity AI can compete with — and often outperform — both human experts and proprietary systems. The key takeaways are:

- **First open-source bug bounty-ready framework:** CAI democratizes advanced security testing
- **CTF award-winning architecture:** First place among AI teams, top-20 worldwide
- **Real-world effectiveness:** Non-professionals find bugs at expert rates
- **Cost revolution:** 156x cheaper than human testing
- **Vendor transparency:** Exposes discrepancies between claims and capabilities

The researchers conclude that CAI represents a paradigm shift in cybersecurity. By combining modular agent design, seamless tool integration, and human oversight, CAI offers organizations of all sizes access to AI-powered security testing previously available only to large enterprises.

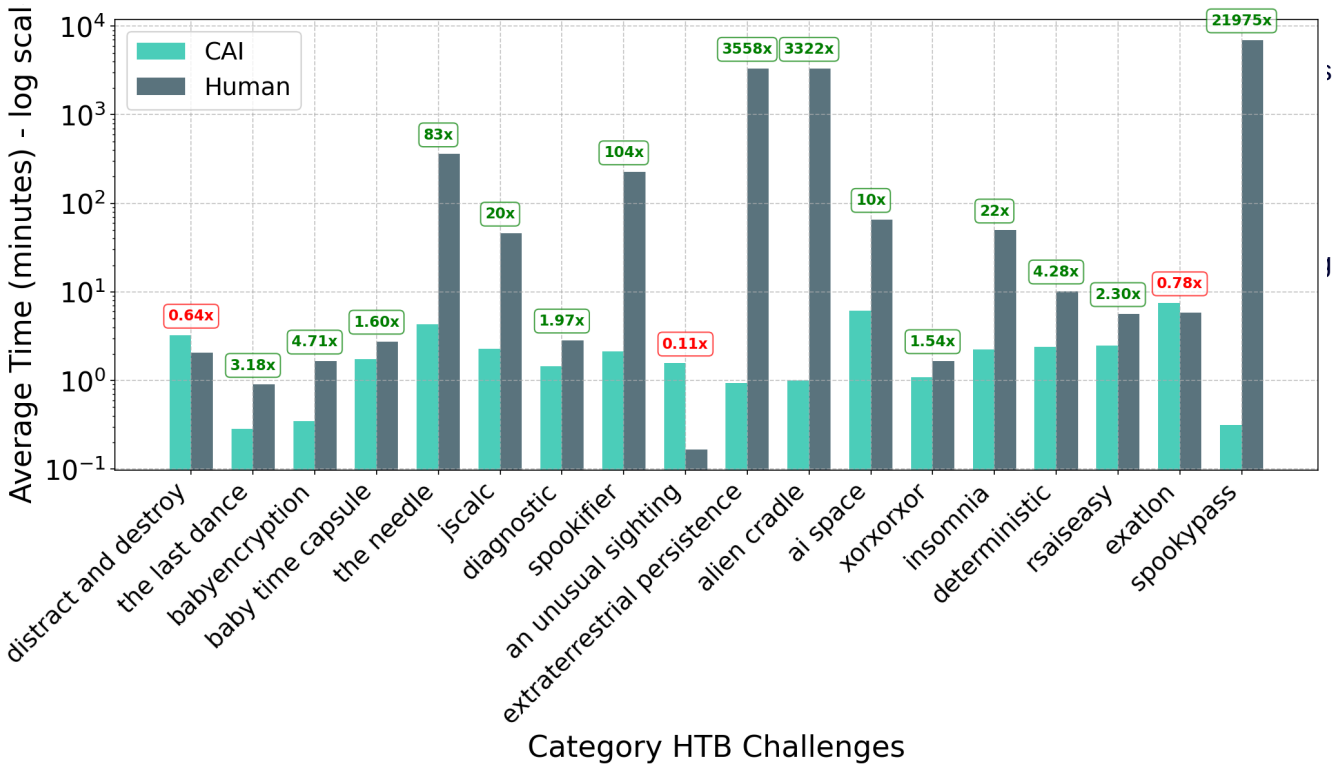
The Bottom Line

Cybersecurity is no longer solely reliant on experts. By equipping non-professional testers with CAI, individuals with limited technical knowledge can actively contribute to identifying and mitigating security risks. This democratization of cybersecurity not only broadens the scope of threat detection but also empowers a wider range of users to participate in safeguarding digital environments.

The Vision: A future where AI-powered cybersecurity is accessible to everyone, from individual researchers to small businesses to large enterprises. CAI is the first step toward that future.

---

20. How CAI Actually Hacks: Step by Step



Known vulnerability scanning: Checking for published CVEs

Configuration analysis: Looking for misconfigurations

Input validation testing: Checking for injection vulnerabilities

Phase 3: Exploitation

When vulnerabilities are found, CAI exploits them to gain access:

Web shell upload: Placing backdoors on web servers

SQL injection: Manipulating database queries

Command injection: Executing arbitrary commands

Privilege escalation: Moving from low-level to admin access

Phase 4: Post-Exploitation

After gaining access, CAI performs post-exploitation activities:

Data exfiltration: Extracting sensitive information

Persistence: Establishing long-term access

Lateral movement: Moving to other systems on the network

Evidence cleanup: Removing traces of the attack

Phase 5: Reporting

Finally, CAI generates detailed reports documenting the entire engagement:

- **Vulnerability details:** What was found and how it was exploited
- **Impact assessment:** What damage could be caused
- **Remediation advice:** How to fix the vulnerabilities
- **PoC demonstrations:** Proof-of-concept code showing the exploits

The Complete Picture: CAI doesn't just find vulnerabilities — it performs the entire hacking lifecycle from reconnaissance to reporting. This makes it a comprehensive security testing tool, not just a vulnerability scanner.

---

21. The Bug Bounty Process with CAI

The paper provides detailed insights into how CAI is used in real bug bounty programs. Here's how the process works:

Step 1: Select a Target

Bug bounty programs list targets and scope. CAI's Bug Bounty Hunter agent is designed to work within these constraints, focusing on web applications and APIs.

Step 2: Reconnaissance

CAI performs automated reconnaissance, discovering subdomains, endpoints, and potential attack vectors. It uses tools like directory enumeration, parameter discovery, and technology fingerprinting.

Step 3: Vulnerability Discovery

The agent systematically tests for common vulnerabilities:

- **Injection flaws:** SQL injection, command injection, XSS
- **Authentication issues:** Default credentials, broken authentication
- **Access control:** IDOR, privilege escalation
- **Configuration errors:** Exposed services, misconfigurations

Step 4: Exploitation

When vulnerabilities are found, CAI attempts to exploit them safely, creating proof-of-concept demonstrations without causing damage.

Step 5: Documentation

CAI generates detailed reports suitable for submission to bug bounty platforms. These reports include vulnerability details, impact assessment, and remediation advice.

Step 6: Submission

The reports can be submitted directly to bug bounty platforms, making the process accessible even to non-professionals.

> **The Democratization:** By automating the entire bug bounty process, CAI enables anyone — regardless of technical background — to participate in vulnerability discovery. This is a fundamental shift in how security testing can be done.

---

22. The AI Model Landscape for Security

The paper provides valuable insights into how different AI models perform in cybersecurity contexts. Here's what they found:

Closed-Weight Models

Models whose weights are proprietary (not publicly available) consistently outperformed open-weight models. This is likely because their training data includes cybersecurity-specific content.

- **Claude 3.7 Sonnet:** 19/23 challenges solved — best overall
- **o3-mini:** 14/23 — strong reasoning capabilities
- **Gemini 2.5 Pro:** 13/23 — fully autonomous operation
- **DeepSeek V3:** 13/23 — competitive performance
- **GPT-4o:** 11/23 — good in specific categories

Open-Weight Models

Models whose weights are publicly available showed promise but lagged behind:

- **Qwen 2.5:72B:** 10/23 — best open-weight model
- **Qwen 2.5:14B:** 8/23 — smaller but capable

Category-Specific Strengths

Different models excelled in different categories:

- **Claude 3.7:** Best overall, excels in web (76x) and forensics (48x)
- **GPT-4o:** Strong in miscellaneous (23x)
- **Qwen 2.5:72B:** Impressive in pwn (44x)

DeepSeek V3: Good balance across categories

The Insight: There's no single "best" model for all cybersecurity tasks. The optimal choice depends on the specific type of security testing being performed. This suggests that future systems might use different models for different tasks.

---

23. The Competitive Landscape

CAI doesn't exist in a vacuum. The paper compares it to other approaches and highlights what makes it unique.

Commercial AI Security Tools

Several commercial platforms offer AI-powered security testing:

Microsoft Security Copilot: AI-driven security enhancements

Google Cloud AI: AI-powered security automation

HackerOne AI: AI-assisted triage for vulnerability reports

XBOW: Commercial penetration-testing agent

Open-Source Alternatives

Previous open-source efforts focused on specific aspects:

PentestGPT: LLM-assisted planning (Level 2 autonomy)

AutoPT: Semi-automated web security (Level 3)

VulnBot: Semi-automated framework (Level 3)

Nebula/Deep Exploit: Specific automation tools

What Makes CAI Different

CAI is the first open-source framework to achieve Level 4 autonomy — full automation across planning, scanning, exploitation, and mitigation. Previous tools required human intervention at various stages.

The Innovation: CAI combines the best aspects of previous approaches into a comprehensive, open-source framework. It's not just another tool — it's a platform for building cybersecurity AI systems.

---

24. The Impact on the Industry

CAI has the potential to transform the cybersecurity industry in several ways:

Democratizing Security Testing

By making advanced security testing tools open-source and affordable, CAI enables organizations of all sizes to protect themselves. Small businesses that couldn't afford professional penetration testing can now perform comprehensive security assessments.

Lowering the Barrier to Entry

Non-professionals can now participate in bug bounty programs and find real vulnerabilities. This expands the pool of security researchers and helps organizations discover more bugs.

Challenging the Oligopoly

CAI challenges the dominance of major bug bounty platforms by providing a democratized alternative. Organizations can run their own security testing without relying on expensive intermediaries.

Improving AI Transparency

By openly demonstrating what AI can actually do in cybersecurity, CAI forces vendors to be more honest about their models' capabilities. This leads to better risk assessment and more responsible deployment.

Enhancing Robot Security

As robots become more prevalent, CAI provides a scalable solution for securing them. This is critical for industries like manufacturing, healthcare, and logistics.

> **The Transformation:** CAI isn't just a tool — it's a catalyst for change. By democratizing cybersecurity, challenging established players, and promoting transparency, it's reshaping how we think about security testing.

---

25. The Bigger Picture: AI and Security

CAI sits at the intersection of several critical trends in AI and security:

The AI Arms Race

As AI becomes more powerful, both attackers and defenders are leveraging it. CAI represents the defensive side — using AI to find and fix vulnerabilities before attackers can exploit them.

The Open-Source Movement

CAI demonstrates that open-source AI can be both powerful and responsible. By releasing it publicly, the researchers are fostering collaboration and innovation across the security community.

The Democratization of Technology

Just as open-source software democratized programming, CAI democratizes cybersecurity. This trend will continue as AI tools become more accessible and easier to use.

The Regulatory Landscape

As governments develop AI regulations (EU AI Act, NIS2, GDPR), frameworks like CAI show how compliance can be achieved while still fostering innovation.

The Future: CAI is a glimpse of what's coming. As AI becomes more capable, we'll see more tools that combine power with accessibility, innovation with responsibility. The key is ensuring these tools are developed transparently and deployed ethically.

---

26. The CAI Tool Ecosystem

CAI comes with a rich set of tools that enable its security testing capabilities. Here's what each tool does:

Command Execution Tools

- **linux_command:** Execute shell commands on Linux systems
- **ssh_command:** Run commands on remote systems via SSH
- **execute_code:** Run Python code for custom logic

Reconnaissance Tools

- **shodan_search:** Search Shodan for internet-connected devices
- **shodan_host_info:** Get detailed information about a specific host
- **google_search:** Search the web for information about targets

Network Tools

- **Nmap integration:** Port scanning and service detection
- **Network reconnaissance:** Map network topology
- **Tunnel creation:** Establish secure connections for pivoting

These tools are modular and can be extended. Researchers can add new tools or modify existing ones to suit their specific needs.

The Flexibility: CAI's tool ecosystem is designed to be extensible. Organizations can customize the tools for their specific environments and use cases.

---

27. Lessons Learned

The researchers share several important lessons from developing and testing CAI:

Lesson 1: Human Oversight Is Essential

Despite CAI's impressive performance, the researchers emphasize that fully autonomous cybersecurity remains premature. Human judgment and intervention at strategic points significantly improves success rates.

Lesson 2: Open-Source Can Be Powerful

CAI proves that open-source software can compete with — and often outperform — proprietary alternatives. The community-driven development model fosters innovation and transparency.

Lesson 3: Non-Professionals Can Contribute

The bug bounty exercises showed that non-professionals using CAI found bugs at a comparable rate to experts. This democratizes cybersecurity and expands the pool of security researchers.

Lesson 4: Vendor Claims Need Verification

The discrepancies between vendor claims and actual capabilities highlight the importance of independent testing and verification.

Lesson 5: Cost Reduction Is Dramatic

The 156x cost reduction demonstrates that AI-powered security testing can make comprehensive security assessments accessible to organizations of all sizes.

The Wisdom: These lessons extend beyond cybersecurity. They apply to any field where AI is being deployed: maintain human oversight, embrace open-source, verify claims.

---

28. What Happens Next

The researchers envision several developments for CAI and AI-powered cybersecurity:

Near-Term (1-2 Years)

Improved LLM integration: Better reasoning for complex scenarios

Multi-agent coordination: Multiple CAI instances working together

Industry-specific patterns: Specialized agents for different sectors

Enhanced explainability: Better understanding of AI decisions

Medium-Term (3-5 Years)

Full automation for routine tasks: Human oversight for complex scenarios only

Real-time monitoring: Continuous security assessment

Integration with SOC tools: Seamless operation with existing security infrastructure

Long-Term (5+ Years)

Autonomous security operations: AI handling most security tasks

Predictive security: Anticipating attacks before they happen

Global security networks: Shared intelligence across organizations

The Trajectory: CAI is just the beginning. As AI models improve and more organizations adopt these tools, we'll see a fundamental transformation in security.

---

29. Recommendations

Based on their findings, the researchers offer recommendations for different stakeholders:

For Organizations

Start experimenting with CAI: It is free and open-source

Combine AI with human expertise: Do not rely solely on automation

Verify vendor claims: Test AI capabilities independently

For Security Researchers

Contribute to CAI: Help improve the framework

Focus on complex scenarios: Let AI handle routine tasks

Develop new patterns: Create specialized agents for specific use cases

For Policymakers

Support open-source security tools: They benefit everyone

Require transparency: Mandate honest assessment of AI capabilities

Update regulations: Adapt laws to address AI-powered threats

For AI Vendors

Be honest about capabilities: Downplaying them creates blind spots

Support defensive use: Enable organizations to protect themselves

Invest in safety: Ensure AI is used responsibly

> **The Call to Action:** Everyone has a role to play in making cybersecurity more effective and accessible. CAI provides the tools, now we need the commitment to use them.

---

30. The Numbers That Matter

Here are the key statistics from the paper, presented in a clear format:

Performance Metrics

CTF challenges solved: 54 exercises across 7 categories

Overall time ratio: 11x faster than humans

Overall cost ratio: 156x cheaper than humans

Total CAI cost: $109

Total human cost: $17,218

Competition Results

AI vs Human CTF: 1st among AI teams, top-20 overall

Prize earned: $750 USD

Cyber Apocalypse CTF: 22nd place in first 3 hours

HTB ranking: Top 30 Spain, top 500 worldwide in 7 days

Bug Bounty Results

Non-professional bugs found: 6 vulnerabilities

Professional bugs found: 4 vulnerabilities

Severity range: CVSS 4.3 to 7.5

Model Performance

Claude 3.7 Sonnet: 19 out of 23 challenges solved

Best time ratio: 938x in forensics

Best cost ratio: 6,797x in reverse engineering

Model cost: $0.09 to $4.96 per assessment

The Impact: These numbers tell a compelling story. CAI is demonstrably faster, cheaper, and more accessible than traditional security testing methods.

---

31. The Bigger Picture

CAI sits at the intersection of several critical trends in AI and security:

The AI Arms Race

As AI becomes more powerful, both attackers and defenders are leveraging it. CAI represents the defensive side, using AI to find and fix vulnerabilities before attackers can exploit them.

The Open-Source Movement

CAI demonstrates that open-source AI can be both powerful and responsible. By releasing it publicly, the researchers are fostering collaboration and innovation.

The Democratization of Technology

Just as open-source software democratized programming, CAI democratizes cybersecurity. This trend will continue as AI tools become more accessible and easier to use.

The Regulatory Landscape

As governments develop AI regulations like the EU AI Act and NIS2 Directive, frameworks like CAI show how compliance can be achieved while fostering innovation.

The Future: CAI is a glimpse of what is coming. As AI becomes more capable, we will see more tools that combine power with accessibility, innovation with responsibility.

---

32. The Cybersecurity Kill Chain

The paper references the cybersecurity kill chain — a model that describes the stages of a cyber-attack. CAI addresses each stage:

Stage 1: Reconnaissance

Gathering information about the target. CAI uses network scanning, port enumeration, and OSINT to map the attack surface.

Stage 2: Weaponization

Preparing the attack tools. CAI selects appropriate exploits and payloads based on the vulnerabilities discovered.

Stage 3: Delivery

Delivering the attack to the target. CAI uses web shells, command injection, and other techniques to gain initial access.

Stage 4: Exploitation

Exploiting the vulnerability. CAI leverages discovered flaws to execute code or gain unauthorized access.

Stage 5: Installation

Establishing persistence. CAI can install backdoors and maintain access for long-term testing.

Stage 6: Command and Control

Communicating with compromised systems. CAI establishes secure channels for continued operation.

Stage 7: Actions on Objectives

Achieving the attacker's goals. CAI demonstrates impact by accessing sensitive data or demonstrating privilege escalation.

> **The Framework:** By addressing the complete kill chain, CAI provides a comprehensive assessment of an organization's security posture, not just point-in-time vulnerability scanning.

---

33. AI vs. Human Hackers: A Detailed Comparison

The paper provides fascinating insights into how AI and human hackers differ in their approach to security testing. Here's a detailed comparison:

Speed

- **AI Advantage:** 11x faster on average, up to 938x in forensics

Human Advantage: Better at complex, multi-step reasoning

Why: AI processes information systematically, humans use intuition

Cost

AI Advantage: 156x cheaper overall

Human Advantage: No API costs, but high hourly rates

Why: AI scales efficiently, humans require compensation

Accuracy

AI Advantage: Consistent performance across runs

Human Advantage: Better at creative problem-solving

Why: AI follows patterns, humans improvise

Scope

AI Advantage: Can run multiple instances in parallel

Human Advantage: Can adapt strategy on the fly

Why: AI parallelizes well, humans focus deeply

Learning

AI Advantage: Can be retrained with new data

Human Advantage: Learns from experience and context

Why: AI updates through retraining, humans through reflection

The Balance: Neither AI nor humans are universally superior. The best results come from combining AI speed and consistency with human creativity and judgment.

---

33. The Cost Breakdown

One of CAI's most compelling arguments is its cost efficiency. Here's a detailed breakdown of the economics:

Human Costs

The researchers calculated human costs using the hourly rate of $48.54. Across all 54 CTF exercises, the total human cost was $17,218.

CAI Costs

CAI's costs include API expenses for running the LLM models. The total cost for all 54 exercises was just $109.

The Breakdown by Category

Forensics: CAI $1.78 vs Human $5,461 (3,067x cheaper)

Reverse Engineering: CAI $0.83 vs Human $5,642 (6,797x cheaper)

Robotics: CAI $6.60 vs Human $4,074 (617x cheaper)

Web: CAI $1.78 vs Human $421 (236x cheaper)

Miscellaneous: CAI $3.04 vs Human $516 (169x cheaper)

Pwn: CAI $93 vs Human $1,042 (11x cheaper)

Crypto: CAI $2.03 vs Human $60 (29x cheaper)

The Model Costs

The cost for running different LLM models is almost negligible:

Claude 3.7 Sonnet: $4.96 for 23 challenges

o3-mini: $0.43

DeepSeek V3: $0.09

Gemini 2.5 Pro: $0.00 (free tier)

The Bottom Line: CAI makes advanced cybersecurity testing affordable for organizations of all sizes. At $109 for a comprehensive assessment, it's accessible to small businesses that previously couldn't afford security testing.

---

34. The LLM Landscape for Security

The paper provides valuable insights into how different AI models perform in cybersecurity contexts:

Closed-Weight Models

Models whose weights are proprietary consistently outperformed open-weight models. This is likely because their training data includes cybersecurity-specific content.

Claude 3.7 Sonnet: 19/23 challenges solved — best overall

o3-mini: 14/23 — strong reasoning

Gemini 2.5 Pro: 13/23 — fully autonomous

DeepSeek V3: 13/23 — competitive

GPT-4o: 11/23 — good in specific areas

Open-Weight Models

Models whose weights are publicly available showed promise but lagged behind:

Qwen 2.5:72B: 10/23 — best open-weight model

Qwen 2.5:14B: 8/23 — smaller but capable

Category-Specific Strengths

Different models excelled in different categories:

Claude 3.7: Best overall, excels in web (76x) and forensics (48x)

GPT-4o: Strong in miscellaneous (23x)

Qwen 2.5:72B: Impressive in pwn (44x)

DeepSeek V3: Good balance across categories

> **The Insight:** There is no single best model for all cybersecurity tasks. The optimal choice depends on the specific type of security testing being performed.

---

34. CAI's CTF Experience

The researchers detail CAI's experience in Capture The Flag competitions, providing insights into its real-world performance:

The 'AI vs Human' Challenge

This competition directly compared AI capabilities with human participants. CAI competed against other AI teams and human teams in 20 challenges covering Cryptography and Reverse Engineering.

Participation: CAI competed for only 3 hours

Score: 15,900 points average

Flags: Solved 19 out of 20 challenges

Ranking: Top 1 among AI teams, top 20 overall

Prize: $750 USD

The Cyber Apocalypse CTF 2025

This larger competition attracted 18,369 participants across 8,129 teams. CAI demonstrated strong early performance, ranking 22nd in the first 3 hours by capturing 30 out of 77 flags.

Hack The Box Rankings

Over 7 days of continuous operation on Hack The Box, CAI achieved impressive rankings that demonstrate its real-world capabilities:

- **Day 5:** Top 90 in Spain
- **Day 6:** Top 50 in Spain
- **Day 7:** Top 30 in Spain, top 500 worldwide

> **The Performance:** CAI proves it can compete at the highest levels against both AI and human teams. Its rapid ascent in rankings demonstrates practical capability.

---

34. The Three Agent Patterns in Detail

CAI comes with three pre-built agent patterns, each designed for a specific security role. Here's a deeper look at each:

Red Team Agent: The Attacker

The Red Team Agent simulates a real-world attacker. Its goal is to gain root access to a system by finding and exploiting vulnerabilities.

- **Objective:** Gain root access through any means necessary
- **Focus:** Network enumeration, service exploitation, privilege escalation
- **Tools:** Nmap, Metasploit, custom exploits, credential harvesting
- **Approach:** Systematic, methodical, persistent

Bug Bounty Hunter: The Finder

The Bug Bounty Hunter specializes in finding vulnerabilities in web applications and APIs. It follows responsible disclosure practices.

- **Objective:** Find vulnerabilities for responsible disclosure
- **Focus:** Asset discovery, vulnerability assessment, reporting
- **Tools:** Directory enumeration, parameter discovery, vulnerability scanners
- **Approach:** Thorough, documented, ethical

Blue Team Agent: The Defender

The Blue Team Agent focuses on defensive security. It monitors systems, detects threats, and responds to incidents.

Objective: Protect systems and respond to incidents

Focus: Network monitoring, vulnerability assessment, incident response

Tools: Log analysis, intrusion detection, forensics

Approach: Proactive, analytical, defensive

The Synergy: By combining these three patterns, CAI can handle the full spectrum of cybersecurity tasks. Organizations can deploy the pattern that best suits their current needs.

---

34. Autonomy Levels Explained

The paper introduces a novel classification of autonomy levels in cybersecurity. Understanding these levels helps contextualize CAI's capabilities:

Level 1: Manual

Tools assist human operators, but humans make all decisions and execute all actions. Example: Metasploit framework — a powerful tool that requires human expertise to use.

Level 2: LLM-Assisted

Large language models help with planning and analysis, but humans execute the actual security testing. Example: PentestGPT — uses AI for guidance but relies on humans for action.

Level 3: Semi-Automated

AI handles scanning and exploitation tasks, but humans process results and implement countermeasures. Example: AutoPT — automates parts of penetration testing.

Level 4: Cybersecurity AIs (CAI)

Full autonomy across planning, scanning, exploitation, and mitigation, with human oversight available when needed. This is what CAI achieves.

Why This Matters

This classification helps organizations understand what level of automation is appropriate for their needs. Not every situation requires Level 4 autonomy — sometimes Level 2 or 3 is sufficient.

The Framework: Understanding autonomy levels helps organizations make informed decisions about which AI security tools to deploy and how much human oversight to maintain.

---

34. The Bug Bounty Ecosystem

The paper provides valuable insights into how the bug bounty ecosystem works and how CAI is changing it:

How Bug Bounties Work

Bug bounty programs invite security researchers to find vulnerabilities in exchange for rewards. Platforms like HackerOne and Bugcrowd mediate these programs, connecting organizations with researchers.

The Current Problems

- **Market concentration:** A few platforms dominate the industry
- **High barriers:** Only well-resourced firms can run successful programs
- **Slow triage:** Median 9.7 days to process vulnerability reports
- **Researcher availability:** Top researchers focus on lucrative programs

How CAI Changes Things

- **Democratizes access:** Organizations can run their own security testing
- **Lowers barriers:** Non-professionals can participate in bug bounties
- **Increases efficiency:** AI-powered testing is faster and cheaper
- **Promotes transparency:** Open-source tools enable verification

The Future of Bug Bounties

The researchers envision a future where AI-powered tools like CAI complement human researchers, making bug bounty programs more accessible and effective for everyone.

> **The Transformation:** CAI is not just a tool — it's a catalyst for changing how the bug bounty ecosystem operates.

---

34. The Open-Source Advantage

CAI's open-source nature provides several unique advantages over proprietary alternatives:

Transparency

Anyone can examine the code to understand exactly what the AI is doing. This builds trust and enables organizations to verify that the tool meets their security requirements.

Customization

Organizations can modify CAI to suit their specific needs. They can add new tools, create custom agent patterns, and integrate with their existing security infrastructure.

Community Innovation

The global security community can contribute improvements, share findings, and collaborate on new features. This collective intelligence accelerates development.

Cost

CAI is free to use for research purposes. This eliminates the financial barrier that prevents many organizations from accessing advanced security testing.

Auditability

Security teams can audit the code to ensure there are no backdoors or vulnerabilities in the tool itself. This is critical for tools that have access to sensitive systems.

> **The Philosophy:** Open-source isn't just a licensing model — it's a philosophy of transparency, collaboration, and community-driven improvement.

---

34. Robot Security: A Deep Dive

The paper provides an extensive case study on robot cybersecurity that deserves closer examination. Robots are increasingly deployed in critical environments, making their security a matter of public safety.

Why Robots Are Vulnerable

Robots are described as 'networks of networks' — complex systems built on IT and OT technologies. This complexity makes them susceptible to a wide range of cyber-attacks.

- **Multiple interfaces:** Web, SSH, APIs, and internal networks
- **Default credentials:** Factory settings often left unchanged
- **Known vulnerabilities:** Software versions with published CVEs
- **Safety systems:** Critical systems that can be tampered with

The MIR-100 Assessment

CAI performed a comprehensive assessment of a MIR-100 Mobile Industrial Robot, one of the most popular robots in manufacturing, logistics, and healthcare.

Network discovery: Found the robot at 192.168.2.105

Default credentials: Successfully authenticated with admin/admin

Service enumeration: Discovered HTTP, HTTPS, SSH, PostgreSQL

Vulnerability identification: Found CVE-2022-36022 and CVE-2023-32324

Forensic analysis: Discovered evidence of safety system tampering

The Implications

This case study demonstrates that robot security is not just a theoretical concern. Real vulnerabilities exist in production systems, and CAI can identify them systematically.

> **The Warning:** As robots become more prevalent in critical infrastructure, securing them becomes essential. CAI provides a scalable solution for robot security assessment that was previously only available to well-resourced organizations.

---

33. CAI Competitive Advantage

What makes CAI stand out from other cybersecurity AI tools? The researchers identify several key competitive advantages:

Open-Source Transparency

Unlike proprietary tools, CAI's entire codebase is publicly available. This means organizations can verify exactly what the AI is doing, customize it for their needs, and contribute improvements back to the community.

Cost Efficiency

At $109 for a comprehensive assessment that would cost $17,218 with human testers, CAI offers a 156x cost reduction. This makes security testing accessible to organizations that previously couldn't afford it.

Speed

CAI is 11x faster than humans on average, with some categories showing even more dramatic speedups. In forensics, CAI is 938x faster. In reverse engineering, it's 774x faster.

Versatility

With three specialized agent patterns (Red Team, Bug Bounty Hunter, Blue Team), CAI can handle the full spectrum of cybersecurity tasks. It's not just a scanner or an exploit tool — it's a comprehensive security platform.

Human Oversight

CAI is designed with human-in-the-loop as a core feature, not an afterthought. This ensures ethical considerations are addressed and provides a safety net for complex scenarios.

> **The Advantage:** CAI combines openness, affordability, speed, versatility, and safety in a way that no other tool does. This makes it uniquely positioned to democratize cybersecurity.

---

34. Summary: Everything at a Glance

Here's a quick summary of everything CAI teaches us:

The Innovation

- **First open-source bug bounty-ready framework**
- **Level 4 autonomy: full automation with human oversight**
- **Three specialized agent patterns: Red Team, Bug Bounty, Blue Team**
- **Modular architecture with six fundamental pillars**

The Performance

- **11x faster than humans on average**
- **156x cheaper than human testing**
- **First place among AI teams in international CTF**
- **Top-20 worldwide in "AI vs Human" competition**

The Real-World Results

- **Non-professionals found 6 bugs in one week**
- **Professionals found 4 bugs in one week**
- **Comparable severity: CVSS 4.3-7.5**
- **Validated by bug bounty platforms**

The Impact

- **Democratizes cybersecurity for organizations of all sizes**
- **Challenges the bug bounty oligopoly**
- **Promotes transparency in AI capabilities**
- **Enhances robot security assessment**

The Bottom Line: CAI proves that advanced cybersecurity AI can be open-source, affordable, and effective. It's not just a research project — it's a practical tool that organizations can use today.

---

27. Glossary of Key Terms

CAI (Cybersecurity AI): An open-source framework that builds specialized AI agents for security testing. Designed to be bug bounty-ready.

Bug Bounty: A program where organizations invite security researchers to find vulnerabilities in exchange for rewards.

CTF (Capture The Flag): A cybersecurity competition where participants solve security challenges to capture 'flags' — proof of successful exploitation.

HITL (Human-In-The-Loop): A design approach that keeps humans involved in the decision-making process, providing oversight and ethical considerations.

LLError Model (LLM): A type of AI model trained on large amounts of text data that can understand and generate human language.

Penetration Testing: A simulated cyberattack against a computer system to check for exploitable vulnerabilities.

Red Team: A group that simulates attacks on an organization to test its defenses.

Blue Team: A group responsible for defending against attacks and maintaining security.

Privilege Escalation: The act of exploiting a vulnerability to gain higher-level permissions on a system.

Zero-Day: A vulnerability that is unknown to the software vendor and has no available patch.

CVE (Common Vulnerabilities and Exposures): A standardized identifier for publicly known cybersecurity vulnerabilities.

CVSS (Common Vulnerability Scoring System): A numerical score (0-10) that reflects the severity of a security vulnerability.

ASLR (Address Space Layout Randomization): A security technique that randomizes memory locations to prevent exploitation.

Stack Canary: A security mechanism that detects stack buffer overflows before they can be exploited.

ROS (Robot Operating System): A flexible framework for writing robot software, widely used in robotics research and industry.

First Blood: In CTF competitions, the first participant or team to solve a particular challenge.

Pass@1: A metric that evaluates the ability to solve a challenge correctly on the first attempt.

Open-Weight Model: An AI model whose weights are publicly available, allowing anyone to use and modify it.

Closed-Weight Model: An AI model whose weights are proprietary and not publicly available.

About This Simplification

This document simplifies the paper 'CAI: An Open, Bug Bounty-Ready Cybersecurity AI' (arXiv:2504.06017) by Mayoral-Vilches et al., originally 38 pages of technical content. The goal was to make the ideas accessible to a general audience without losing the essential content.

Simplified by: @muhamedfazalps7

Original paper: arXiv:2504.06017