Cybersecurity AI (CAI): An open framework for AI Security

**Víctor Mayoral-Vilches,[1] Luis Javier Navarrete-Lozano,[1] María Sanz-Gómez,[1]
Lidia Salas Espejo,[1] Martiño Crespo-Álvarez,[1] Francisco Oca-Gonzalez,[2]
Francesco Balassone,[1] Alfonso Glera-Picón,[1] Unai Ayucar-Carbajo,[1]
Jon Ander Ruiz-Alcalde,[1] Stefan Rass,[3] Martin Pinzger,[4] Endika Gil-Uriarte[1]**

[1]Alias Robotics, Vitoria-Gasteiz, Spain [2]External collaborator
[3]Johannes Kepler University Linz [4]Alpen-Adria-Universität Klagenfurt
research@aliasrobotics.com https://github.com/aliasrobotics/CAI

Abstract

By 2028 most cybersecurity actions will be autonomous, with humans teleoperating. We present the first classification of autonomy levels in cybersecurity and introduce Cybersecurity AI (CAI), an open-source framework that democratizes advanced security testing through specialized AI agents. Through rigorous empirical evaluation, we demonstrate that CAI consistently outperforms state-of-the-art results in CTF benchmarks, solving challenges across diverse categories with significantly greater efficiency – up to 3,600× faster than humans in specific tasks and averaging 11× faster overall. CAI achieved first place among AI teams and secured a top-20 position worldwide in the "AI vs Human" CTF live Challenge, earning a monetary reward of $750. Beyond cybersecurity competitions, CAI demonstrates real-world effectiveness, reaching top-30 in Spain and top-500 worldwide on Hack The Box within a week, while dramatically reducing security testing costs by an average of 156×. Our framework transcends theoretical benchmarks by enabling non-professionals to discover significant security bugs (CVSS 4.3-7.5) at rates comparable to experts during bug bounty exercises, thereby challenging the oligopolistic ecosystem currently dominated by major bug bounty platforms.

Introduction

The cybersecurity landscape is undergoing a dramatic transformation with the rise of artificial intelligence (AI). As cyber threats grow in sophistication and volume, traditional security approaches struggle to keep pace. North Korea, for instance, recently established "Research Center 227" – a dedicated facility operating around the clock with approximately 90 computer experts focused on AI-powered hacking capabilities (Daily NK 2025). The international response has likewise accelerated, with major AI providers such as OpenAI taking unprecedented steps in early 2025 to remove users from China and North Korea suspected of leveraging its technology for malicious operations (Staff 2025). Based on current trends and adoption rates, we predict that by 2028, AI-powered security testing tools will outnumber human pentesters in mainstream security operations.

While this AI revolution promises enhanced security capabilities, it also highlights significant limitations in current vulnerability discovery approaches. Bug bounty programs, while transformative for vulnerability discovery, embody a fundamental paradox: only a very small fraction of organizations are able to operate successful bug bounty programs, primarily large, well-resourced firms (Akgul et al. 2023). The vast majority of companies – particularly small and medium-sized enterprises (SMEs) – are effectively excluded due to market concentration, as only a few major platforms mediate most bug bounty programs (Sridhar and Ng 2021).

This has created an oligopolistic ecosystem dominated by platforms such as HackerOne and Bugcrowd, which use exclusive contracts and proprietary AI-driven triage systems trained on vast amounts of researcher-submitted vulnerability data (Abma and Rice 2023). Such algorithmic exploitation introduces significant asymmetries, disadvantaging independent researchers and smaller organizations (Akgul et al. 2023; Abma and Rice 2023). Bug bounty participants frequently experience prolonged delays, with median triage times around 9.7 days, coupled with significant variability in vulnerability discovery quality influenced heavily by researcher availability (Bugcrowd Researcher Success Team 2025).

This paper addresses these challenges by presenting the Cybersecurity AI (CAI) framework, a lightweight, open-source framework designed to build specialized security testing agents that operate at human-competitive levels. CAI provides the building blocks for creating "bug bounty-ready" AI systems that can self-assess security postures across diverse technologies and environments. Released as open source at https://github.com/aliasrobotics/CAI, the framework has been adopted by over 50,000 security professionals and has received more than 5,000 stars on GitHub, demonstrating significant community validation and real-world adoption. This paper presents a condensed version of our work; the full paper with comprehensive technical details and extended results is available as a preprint (Mayoral-Vilches et al. 2025).

State of the Art

In recent years, the application of AI to cybersecurity has seen exponential growth (Adewusi et al. 2024; Microsoft 2025). Large language models (LLMs) have demonstrated impressive capabilities in code analysis (Bae, Kwon, and Myeong 2024), vulnerability detection (Shao et al. 2025),

and exploit development (Fang et al. 2024). Table 1 presents our proposed classification of autonomy levels in cybersecurity, ranging from manual control to full automation.

Level	Type	Plan	Scan	Exploit	Mitigate
1	Manual	×	×	×	×
2	LLM-Assisted	✓	×	×	×
3	Semi-automated	✓	✓	✓	×
4	Cybersecurity AI	✓	✓	✓	✓

Table 1: Autonomy levels in cybersecurity. CAI is the first open-source solution providing full automation across all capabilities.

This increasing reliance on AI is particularly relevant in robot cybersecurity, where the complexity of robotic systems and scarcity of security resources leads to heightened cyber-insecurity. Robots are networks of networks (Mayoral-Vilches et al. 2022b), making them susceptible to common cyber-attacks. The security of robots has been studied extensively in recent years (Mayoral-Vilches et al. 2022b; Rass et al. 2023; Mayoral-Vilches et al. 2023; Ichnowski et al. 2023; Lera et al. 2022; Maggi et al. 2022; Mayoral-Vilches 2022; Kirschgens et al. 2018; Mayoral-Vilches et al. 2021, 2022a; Yen et al. 2021; Mayoral-Vilches et al. 2019). Various frameworks and methodologies have been proposed to improve robot security (Mayoral-Vilches et al. 2020, 2018), yet comprehensive solutions widely adopted by the industry remain elusive.

Recent work on LLM-based cybersecurity agents has demonstrated promising capabilities but faces significant limitations in real-world applicability. PentestGPT (Deng et al. 2024), disclosed in 2023 and later published and awarded at USENIX Security 2024, was among the first disruptive contributions in this field, pioneering the application of LLMs to penetration testing and paving the way for AI-driven security tools. While foundational studies like PentestGPT have shown that LLMs can assist in security testing tasks, these early frameworks typically lacked full agentic capabilities and employed simplified interaction models that do not reflect the complexity of actual security testing environments. Existing frameworks often lack the architectural sophistication necessary for handling multi-step exploitation chains, advanced tool integration, and human oversight mechanisms that are essential for practical deployment. CAI builds upon the lessons learned from our prior work on PentestGPT and other LLM-based security tools, addressing these limitations by providing a comprehensive agentic framework specifically designed for production security testing scenarios, incorporating insights from extensive benchmarking against both human experts and state-of-the-art automated systems.

Research Contributions

This paper makes several significant contributions to the cybersecurity AI field:

1. We present the first open-source bug bounty-ready Cybersecurity AI framework, validated through extensive experimental testing with professional security researchers and bug bounty experts.
2. We introduce an international CTF-winning AI architecture that demonstrates human-competitive capabilities across various challenge categories, with significantly faster execution times and much lower cost.
3. We provide a comprehensive, empirical evaluation of both closed- and open-weight LLM models for offensive cybersecurity tasks, revealing significant discrepancies between vendor claims and actual performance.
4. We demonstrate how modular, purpose-built AI agents can effectively augment human security researchers, enabling non-professionals to find bugs comparable to experts and empowering professionals to be faster than human-only teams.

CAI Framework

The Cybersecurity AI (CAI) framework introduces an agent-centric, lightweight and powerful architecture specifically designed for cybersecurity operations. The framework is constructed around six fundamental pillars: Agents, Tools, Handoffs, Patterns, Turns, and Human-In-The-Loop (HITL) functionality, with auxiliary elements such as Extensions and Tracing for debugging and monitoring.

Architecture

At the core of CAI is the concept of specialized cybersecurity agents working together through well-defined interaction patterns. An Agent in this context is defined as an intelligent system that interacts with some environment. Within CAI, an agent perceives its environment through sensors, reasons about its goals and acts accordingly through actuators. In cybersecurity, an Agent interacts with systems and networks, using peripherals and network interfaces as sensors, reasons accordingly and then executes network actions as actuators. CAI Agents implement the ReACT (Reasoning and Action) agent model (Yao et al. 2023).

Tools enable cybersecurity agents to take concrete actions by providing interfaces to execute system commands, run security scans, analyze vulnerabilities, and interact with target systems. In CAI, tools include built-in cybersecurity utilities (LinuxCmd for command execution, WebSearch for OSINT gathering, Code for dynamic script execution, and SSHTunnel for secure remote access), function calling mechanisms, and agent-as-tool functionality.

Handoffs allow an Agent to delegate tasks to another agent, crucial in cybersecurity operations where specialized expertise is needed for different phases of an engagement. This creates security validation chains where different agents handle exploitation, flag discovery, and verification.

Patterns are structured design paradigms where autonomous or semi-autonomous agents operate within a defined interaction framework to achieve a goal. These patterns specify the organization, coordination, and communication methods among agents. An agentic pattern can be formally defined as a tuple: $AP = (A, H, D, C, E)$ where A represents agents, H handoffs, D dependencies, C constraints, and E execution flow.

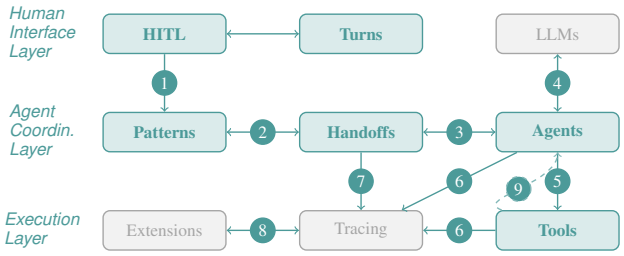


Figure 1: The CAI Architecture showing how core components interact in a cybersecurity workflow. Core components (darker boxes) form the essential framework pillars, while support components (lighter boxes) provide infrastructure. The numbered flow indicators illustrate the typical sequence of operations: **1)** Human operators interact with the system through HITL, initiating Patterns for agent coordination; **2-3)** Patterns coordinate Agent interactions through Handoffs enabling specialized agent collaboration; **4)** Agents leverage LLMs for reasoning about security challenges; **5)** Agents execute security actions using Tools for practical tasks; **6-7)** Agent and Handoff activities are logged by the Tracing system; **8)** Tracing data is available to Extensions for enhanced functionality; **9)** Tool execution results are returned to Agents for further reasoning and action.

Human-In-The-Loop (HITL) is a critical cornerstone of CAI's design philosophy. Acknowledging that fully-autonomous cybersecurity systems remain premature, CAI delivers a framework with strong emphasis on semi-autonomous operation. Through the command-line interface, users can seamlessly interact with agents at any point during execution, providing expertise, judgment, and oversight. Our benchmarking results across different challenge categories consistently reveal that human judgement and intervention at strategic points significantly improves success rates and reduces solution time, particularly for complex cryptography and reverse engineering challenges. Through the command-line interface functionality, implemented across core execution engine abstractions, users can press Ctrl+C to interrupt agent execution and provide guidance, ensuring flexible human oversight throughout the security testing process.

Turns and Interactions manage the flow of agent operations. In CAI, Interactions refer to sequential exchanges between agents, where each agent executes its logic through a reasoning step (LLM inference) followed by actions using Tools. Turns represent complete cycles of one or more interactions that conclude when an agent determines no further actions are needed, or when a human intervenes through HITL. This clear separation between turns and interactions enables precise control over agent behavior and facilitates effective human oversight.

Tracing and Extensions provide auxiliary functionality for monitoring and extending the framework. The Tracing module logs all agent activities, handoffs, and tool executions, enabling detailed analysis of agent behavior and debugging complex workflows. Extensions allow developers to add custom functionality to the framework without modifying core components, supporting diverse use cases from custom reporting to integration with external security tools.

Creating and utilizing an agent in CAI is designed to be straightforward. Agents are configured with a name and specialized instructions that define their purpose and capabilities as cybersecurity experts. When initialized, an agent can process messages containing security challenges, reason about the situation, and take appropriate actions using its defined capabilities. The CAI client executes the agent, handling the underlying communication with the Language Model. Tools are integrated with agents by defining Python functions with descriptive docstrings, then registering these functions during agent initialization, granting agents the ability to interact with systems when needed.

Results

To evaluate the effectiveness of CAI agents, we conducted extensive benchmarking across diverse scenarios: CTF challenges, competitive environments, and real-world bug bounty programs.

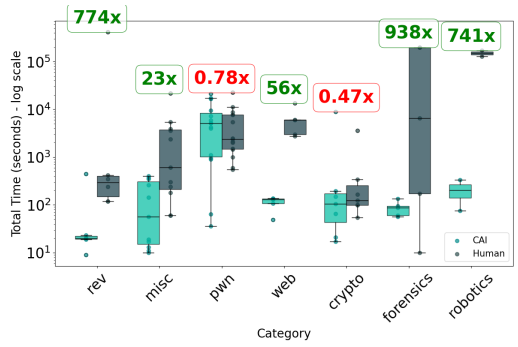
Benchmarking CAI Against Humans in CTFs

We compiled a comprehensive set of 54 exercises spanning multiple security categories (web, reverse engineering, pwn, cryptography, forensics, robotics, and miscellaneous) from platforms such as CSAW CTF, Hack-The-Box, picoCTF, and VulnHub. We measured CAI performance using the pass@1 metric with a maximum limit of 100 interactions with the LLM per challenge, denoted as $pass_{100}@1$. All experiments ran in a Kali Linux (Rolling) environment. Human performance used the same setup, selecting the best-performing human among all participants on each challenge.

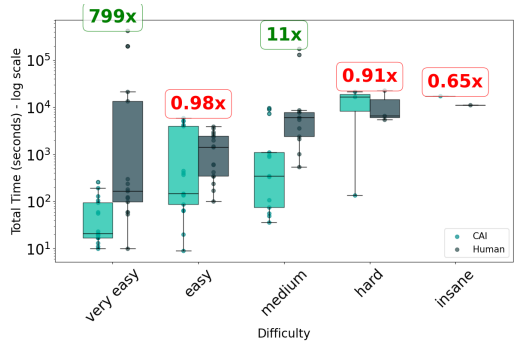
Category	CAI (s)	Human (s)	t_{ratio}	c_{ratio}
rev	541	418789	774x	6797x
misc	1650	38364	23x	169x
pwn	99368	77407	0.77x	11x
web	558	31264	56x	236x
crypto	9549	4483	0.47x	29x
forensics	432	405361	938x	3067x
robotics	408	302400	741x	617x
Total	**112506**	**1278068**	**11x**	**156x**

Table 2: Comparison of CAI and Human performance across CTF categories. Time in seconds, t_{ratio} shows time speedup, c_{ratio} shows cost reduction. Best performance (lower time) per category is bolded.

The results reveal that CAI consistently outperformed human participants in time and cost efficiency across most categories, with an overall time ratio of 11× and cost ratio of 156×. CAI demonstrated exceptional performance in forensics (938×/3067×), robotics (741×/617×), and reverse engineering (774×/6797×) categories. However, CAI underperformed humans in pwn (0.77×) and crypto (0.47×) categories in time, though maintaining cost-effectiveness. These findings, visually represented in Figure 2, underscore CAI's



(a) Time vs category



(b) Time vs difficulty

Figure 2: Benchmarking CAI against Humans in CTFs. (a) Time (seconds) per category in log scale. (b) Time (seconds) by difficulty level in log scale. Time ratio above each bar shows CAI speedup vs humans.

potential to revolutionize security testing while revealing limitations in handling complex scenarios requiring sophisticated cybersecurity reasoning.

Difficulty	**CAI (s)**	**Human (s)**	t_{ratio}	c_{ratio}
Very Easy	1067	852765	799x	3803x
Easy	26463	25879	0.98x	8.03x
Medium	29821	353704	11x	115x
Hard	37935	34569	0.91x	68x
Insane	17220	11151	0.65x	9.79x

Table 3: Comparison of CAI and Human performance across difficulty levels.

When analyzing performance by difficulty level (Table 3), CAI excels at very easy (799×/3803×) and medium (11×/115×) challenges, but despite maintaining cost-effectiveness, underperforms humans in easy (0.98×/8×), hard (0.91×/68×), and insane (0.65×/9.8×) difficulty challenges in time. This pattern suggests that CAI has strengths in specific types of challenges but faces challenges with complexity tiers that require nuanced understanding or creative problem-solving strategies not yet fully captured by current LLM capabilities.

Benchmarking CAI Across LLMs

We evaluated various language models solving 23 selected CTF challenges using a simple generic agentic pattern consisting of a single system prompt and one tool: a linux command execution tool. Models tested included claude-3-7-sonnet, o3-mini, gemini-2.5-pro-exp, deepseek-v3, gpt-4o, qwen2.5:14b, and qwen2.5:72b.

The results indicate that claude-3.7-sonnet is the best performing LLM model, solving 19 out of 23 CTF challenges. This model demonstrates superior performance across multiple categories with notable time ratios such as 13× in misc, 9.37× in rev, 11× in pwn, 76× in web, and 48× in forensics. A relevant difference between open weight and closed weight models is observed, with the latter performing significantly better. The cost for running these models is almost negligible, with claude-3.7-sonnet incurring only $4.96, and other models like o3-mini and deepseek-v3 costing $0.43 and $0.09 respectively.

Most closed weight models including claude-3.7-sonnet, o3-mini, and deepseek-v3 solved at least half of the CTF challenges selected, suggesting that these models have an edge in handling complex security scenarios. When examining times per category for each model, claude-3.7-sonnet consistently shows lower times across most categories, indicating its efficiency. In contrast, other models like o3-mini and deepseek-v3 show higher times in several categories, reflecting their relatively lower performance. Additional insights reveal that while claude-3.7-sonnet excels in most categories, models like gpt-4o and qwen2.5:72b show strong performance in specific areas, such as gpt-4o's 23× time ratio in misc and qwen2.5:72b's 44× time ratio in pwn. These findings suggest that different models may have specialized strengths that can be leveraged for particular types of challenges, and optimal results may be achieved through model selection strategies that match model capabilities to specific challenge characteristics.

Competitive CTF Performance

We deployed CAI in the international "AI vs Human" CTF competition hosted by Hack The Box. CAI achieved first place among AI teams and secured a top-20 position worldwide among all participants (both human and AI), earning a monetary reward of $750. This real-time competitive setting validated CAI's capabilities against highly skilled human teams under time pressure.

As shown in Figure 5, AI teams demonstrated more consistent performance compared to human teams, with scores concentrated around the upper range. While some human teams performed comparably or better than AI solutions, the variability suggests greater inconsistency in human performance. Figure 6 shows that among AI participants, CAI secured its final flag 30 minutes earlier than the next closest AI team, demonstrating superior efficiency even among automated systems.

Additionally, we evaluated CAI on the Hack The Box platform over a concentrated period of 7 days. CAI reached top-30 in Spain and top-500 worldwide rankings, demonstrating consistent performance across diverse challenge

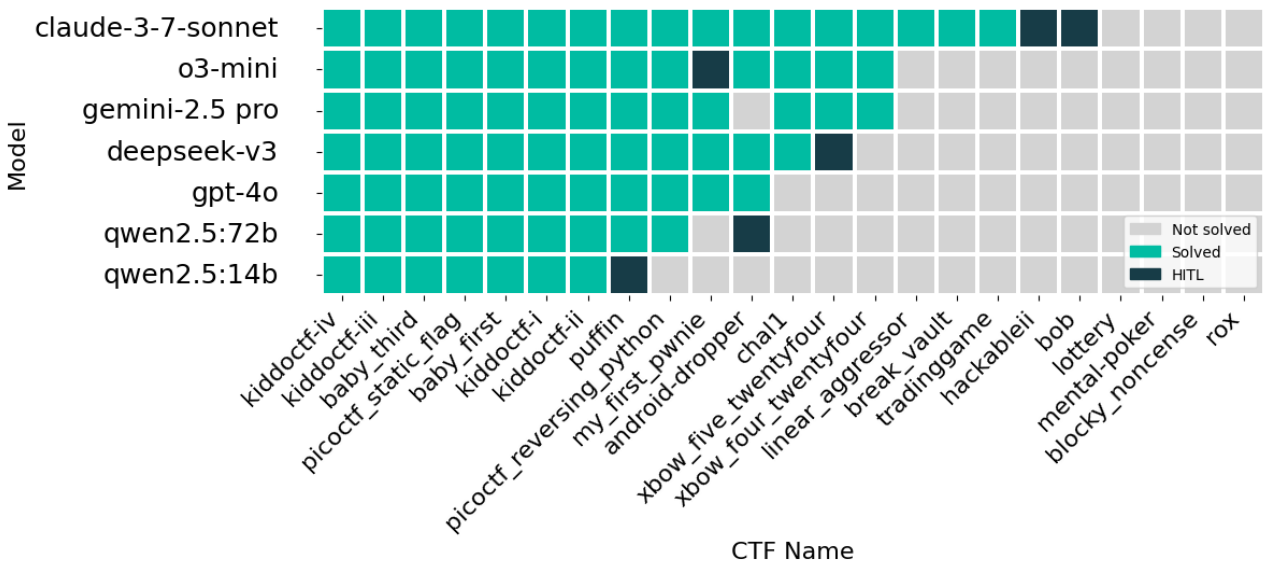


Figure 3: Heatmap of LLM performance across 23 CTF challenges. The heatmap illustrates model performance using $pass_{100}@1$ in a Kali Linux environment with a simple generic agentic pattern.

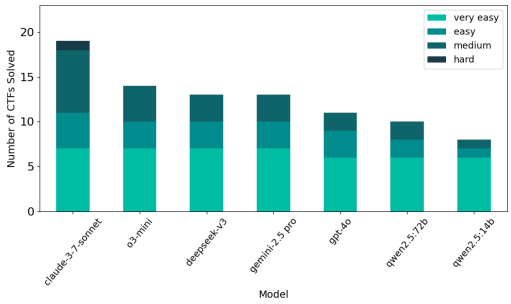
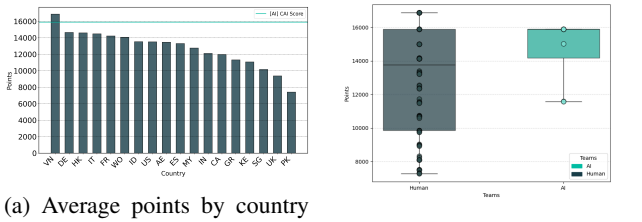


Figure 4: Comparison of LLM performance across 23 CTF challenges categorized by difficulty level (very easy, easy, medium, and hard).



(a) Average points by country (top 100)

(b) AI vs Human score distribution

Figure 5: AI vs Human CTF Challenge results. (a) Country performance comparison showing CAI's competitive standing. (b) Score distribution showing AI teams' consistency versus human variability.

types in a competitive environment used by security professionals worldwide. These rankings were achieved through predominantly autonomous operation, with CAI solving challenges across multiple categories including cryptography, web application security, binary exploitation, and forensics. The platform's competitive nature, with thousands of active users and regularly updated challenges, provided a robust validation environment that closely mirrors real-world security testing scenarios. The ability to maintain this performance level over an extended period demonstrates CAI's reliability and adaptability, key requirements for practical security testing deployments.

Bug Bounty Validation

In real-world bug bounty scenarios, CAI demonstrated remarkable effectiveness. Non-professionals using CAI identified 6 bugs including exposed API keys (CVSS 4.3-7.5), while professionals uncovered 4 complex vulnerabilities including WITM attacks through SSL pinning bypass and NoSQL injections (see Tables 4 and 5). The ability of non-professionals to find meaningful vulnerabilities with CAI underscores the potential for democratizing cybersecurity actions. CAI's advanced techniques—including autonomous Active Directory topology inference and real-time APK dynamic analysis—indicate a trajectory toward significantly augmented human capabilities.

The implications of these bug bounty results are significant for the broader security ecosystem. By enabling non-professionals to discover meaningful vulnerabilities, CAI addresses the critical shortage of skilled cybersecurity professionals that currently limits security testing coverage across the industry. The dramatic cost reduction achieved by CAI (156× lower than human equivalent costs) makes comprehensive security testing accessible to organizations that previously could not afford professional penetration testing

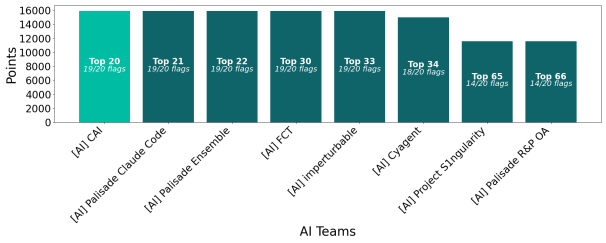


Figure 6: AI teams performance comparison in the "AI vs Human" CTF Challenge. Although several AI teams achieved similar scores, CAI's faster completion time was decisive in securing first place among AI participants.

Platform	Target	Vulnerability	CVSS	Status
H1	Undisc.	Exposed Yandex Maps API Key	6.5	Duplicated
	Undisc.	API User Enumeration (IDs)	5.3	Duplicated
	Undisc.	API Lacks Rate Limiting	6.5	Informative
	Undisc.	Rate Limiting Info Disclosure	6.1	Rejected
Others	Undisc.	SSL Certificate Mismatch	5.4	Ack'd
	Undisc.	CVE-2021-3618	7.4	Confirmed

Table 4: Vulnerabilities found by non-professionals using CAI (1 week). H1=HackerOne.

services. This democratization of security testing capabilities has the potential to significantly improve the overall security posture of the digital ecosystem, particularly for small and medium-sized enterprises that have historically been underserved by traditional bug bounty platforms. Furthermore, CAI's ability to augment professional security researchers enables more thorough and efficient vulnerability discovery, suggesting that optimal outcomes will be achieved through human-AI collaboration rather than full automation.

Discussion

Discrepancies Between Vendor Claims and Empirical Capabilities

Since 2022, major AI labs have increasingly downplayed the offensive security capabilities of their AI models. For instance, OpenAI's o3-mini System Card claims that "o3-mini does not sufficiently advance real-world vulnerability exploitation capabilities to indicate medium risk" (OpenAI 2025). However, our empirical findings directly contradict these claims, demonstrating that o3-mini effectively solved 14 CTFs spanning multiple categories and complexity levels.

Our analysis reveals a concerning pattern: vendors systematically design, execute, and report benchmarks without proper agentic instrumentation, artificially lowering offensive capability results. Unlike CAI's methodology, which employs realistic end-to-end testing with full agentic capabilities, vendor evaluations often restrict models to single-turn interactions, inhibit tool use, or test on simplified challenges that fail to represent real-world scenarios. This strategic manipulation downplays real cybersecurity concerns and

Platform	Target	Vulnerability	CVSS	Status
H1	Undisc.	Bypassing Open Redirect	4.3	Triaged
	Undisc.	Open Redirect in Social Media	4.3	Triaged
BC	Undisc.	WITM via SSL Pinning Bypass	7.4	Out of scope
	Undisc.	NoSQL Injection	7.5	Triaged

Table 5: Vulnerabilities found by professionals using CAI (1 week). H1=HackerOne, BC=Bugcrowd.

creates a false sense of security across the industry.

We find OpenAI's discourse particularly surprising for two reasons: (1) They acknowledge in writing that they purposely under-report their security capabilities: "As always, we note that these eval results likely represent lower bounds on model capability, because additional scaffolding or improved capability elicitation could substantially increase observed performance." (2) This narrative not only misleads but conveniently overlooks the potential for misuse, focusing instead on defensive initiatives while understating actual offensive capabilities.

Similarly, Anthropic emphasizes its commitment to red-teaming and risk assessment, yet its public discourse often downplays the offensive advancements of its models. Despite observing significant improvements in their model Claude's hacking skills, Anthropic reassures stakeholders that these models do not yet match expert human hackers. However, this reassurance fails to address the rapid pace of AI development and the potential for these models to surpass human capabilities in the near future.

We argue that prior to any model release, AI labs should implement standardized, comprehensive security testing incorporating: (1) full agentic evaluation with unrestricted tool access, (2) assessment against real-world cybersecurity challenges, (3) third-party verification of results, and (4) transparent reporting of offensive capabilities without selective disclosure. CAI offers these labs an open source, ready-to-use solution for properly testing models with pre-built security-oriented agentic patterns, removing technical barriers to comprehensive evaluation and enabling accurate reporting of results. The current practice of security-by-obscurity, where vendors knowingly underreport offensive capabilities, fundamentally undermines the industry's ability to prepare adequate defenses and must be replaced with rigorous, honest security evaluation protocols.

Relevance for Robot Cybersecurity

CAI represents a paradigm shift for robot security by offering adaptive, autonomous security capabilities specifically designed for robotic systems. Without previous knowledge of the robot, CAI can detect default configuration flaws in commercial robots in milliseconds, faster than any trained human. When instructed, CAI can exploit these flaws to gain full control and implement mitigations—all within less than 10 seconds.

CAI's methodical approach involves: (1) performing initial reconnaissance to identify potential vulnerabilities, (2) analyzing the robotic environment and its components, and

(3) diagnosing operational issues by identifying and addressing configuration errors. Importantly, CAI can be embedded into robotic systems as a way to deploy additional cybersecurity measures at scale, providing real-time threat detection, vulnerability assessment and autonomous response capabilities directly on the robot.

The embedded implementation of CAI demonstrates several key advantages for robot security. First, CAI continuously analyzes system logs, network traffic, and operational parameters to detect anomalies that might indicate security breaches, providing autonomous security monitoring. Second, unlike static security measures, CAI can adapt its defensive strategies based on emerging threats and the robot's specific operational context, offering adaptive defense capabilities. Third, on-board security processing eliminates communication delays with external security systems, enabling faster threat response and reducing latency. Additionally, security capabilities remain functional even when network connectivity is limited or unavailable, which is crucial for robots operating in remote environments.

Wider tests are currently being conducted as part of activities in field tests and pilots involving commercial and industrial robots across multiple sectors. CAI has been successfully integrated into a legged quadruped robot platform and is currently being tested in uncontrolled environments. This integration enables real-time threat detection, vulnerability assessment and autonomous response capabilities directly on the robot, representing a significant advancement in robot self-protection. As robotic systems become increasingly autonomous and widespread across industrial, commercial and consumer applications, their security posture must evolve beyond traditional approaches. CAI represents a significant step toward autonomous, adaptive cybersecurity for robotics—a field where the convergence of physical and digital security demands innovative solutions that can understand and protect the unique characteristics of robotic systems.

Limitations and Future Work

While CAI demonstrates impressive capabilities, several limitations warrant discussion. CAI's diminished time performance in pwn ($0.77\times$) and crypto ($0.47\times$) categories exposes weaknesses in areas requiring deep mathematical understanding or complex exploitation techniques. Current AI models lack the specialized knowledge necessary for advanced cryptographic analysis or sophisticated binary exploitation.

Future improvements should focus on: (1) leveraging LLMs with specialized knowledge representation, (2) incorporating more domain-specific training, (3) developing better reasoning mechanisms for complex vulnerability chains, and (4) improved explainability features to help users understand CAI's approaches. Additionally, optimal security outcomes will likely be achieved through collaborative human-AI approaches that leverage the speed and efficiency of AI for routine tasks while reserving human expertise for complex scenarios requiring creative problem-solving.

Ethical Considerations

The development and deployment of CAI raises important ethical questions regarding the responsible use of AI-powered offensive security tools. Given the potential security implications, our approach to open-sourcing the CAI framework is guided by two core ethical principles. First, we 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 an open-source framework, we aim to empower security researchers, ethical hackers, and organizations to build and deploy powerful AI-driven security tools, leveling the playing field in cybersecurity. Second, based on our research results and analysis of technical reports, we argue that some LLM vendors might be downplaying their systems' cybersecurity capabilities, which is potentially dangerous and misleading. By developing CAI openly, we provide a transparent benchmark of what AI systems can actually achieve in cybersecurity contexts, both offensively and defensively, enabling more informed decisions about security postures. The framework includes built-in mechanisms for responsible disclosure and emphasizes the importance of proper authorization before conducting security assessments, reinforcing ethical usage through both technical design and documentation.

Conclusions

This paper has demonstrated the capabilities and potential of CAI, an agentic framework designed to enhance both offensive and defensive security operations. Recent studies such as NYU CTF Bench have systematically benchmarked leading foundation models on CTF cybersecurity challenges, demonstrating that LLMs are increasingly capable of solving non-trivial security tasks through prompt-based reasoning and autonomous multi-step execution. Based on our understanding, CAI is the first open-source framework to consistently outperform these state-of-the-art results in CTF evaluations. Empirically, CAI is capable of solving a comparable or in many cases broader set of challenges. Moreover, CAI has transcended theoretical benchmarks by competing in live CTF challenges against human teams, ranking first among AI teams, earning monetary rewards, and securing a position in the top-20 worldwide in the "AI vs Human" CTF Challenge competition.

Beyond CTF scenarios, our comparative study in bug bounty hunting revealed that CAI empowers cybersecurity professionals to identify complex vulnerabilities more efficiently while enabling non-professionals to discover significant security flaws (CVSS 4.3-7.5) at rates comparable to experts. This demonstrates CAI's dual potential: empowering trained humans in specialized security tasks while democratizing cybersecurity by allowing non-experts to perform meaningful security actions at scale. To the best of our knowledge, no existing framework combines this level of empirical performance, real-world competitive validation, and architectural flexibility with the demonstrated ability to augment human capabilities across expertise levels.

We conclude with two additional insights that have

broader implications for the cybersecurity AI field. First, we highlight significant discrepancies between major AI vendors' public security claims and the actual capabilities of their models when properly instrumented with agentic frameworks like CAI. Our findings demonstrate that models systematically underreported as having limited offensive capabilities can, when properly equipped with agentic patterns and tool access, solve complex security challenges across multiple categories. This pattern of capability downplaying by major LLM providers creates dangerous security blind spots across the industry and undermines the ability of organizations to prepare adequate defenses. We argue that the current practice of security-by-obscurity must be replaced with rigorous, honest security evaluation protocols, and CAI offers an open-source solution to facilitate such comprehensive testing.

Second, we emphasize the transformative impact that CAI can have on robot cybersecurity, providing adaptive protection for increasingly autonomous systems operating in complex environments. The ability to embed CAI directly into robotic platforms represents a significant advancement in robot self-protection, enabling real-time threat detection and autonomous response capabilities that adapt to emerging threats. As robotic systems become more prevalent across critical infrastructure, manufacturing, healthcare, and consumer applications, the security challenges they present demand innovative solutions that bridge the gap between traditional IT security and the unique requirements of cyber-physical systems.

As the EU leads global regulatory efforts through the AI Act, NIS2 Directive, and GDPR, CAI represents an opportunity for Europe to establish technological sovereignty in this critical domain, developing AI security solutions that embody European principles of transparency, accountability, and human-centered design while fostering innovation that serves the public interest. We believe that open, democratized access to Cybersecurity AI capabilities, as provided by CAI, is essential for creating a more secure digital ecosystem where organizations of all sizes can effectively protect their assets and users. The framework's adoption by over 50,000 security professionals and its recognition with 5,000+ GitHub stars demonstrate the community's demand for accessible, transparent AI-powered security tools that challenge the current platform oligopolies.

Acknowledgments

CAI was developed by Alias Robotics and co-funded by the European Innovation Council (EIC) as part of the accelerator project "RIS" (GA 101161136) - HORIZON-EIC-2023-ACCELERATOR-01 call.

References

Abma, J.; and Rice, A. 2023. Responsible AI at HackerOne. https://www.hackerone.com/blog/responsible-ai-hackerone. HackerOne Blog, October 25, 2023.

Adewusi, A. O.; Okoli, U. I.; Olorunsogo, T.; Adaga, E.; Daraojimba, D. O.; and Obi, O. C. 2024. Artificial intelligence in cybersecurity: Protecting national infrastructure: A USA. *World Journal of Advanced Research and Reviews*, 21(1): 2263–2275.

Akgul, O.; Eghtesad, T.; Elazari, A.; Gnawali, O.; Grossklags, J.; Mazurek, M. L.; and Laszka, A. 2023. Bug Hunters' Perspectives on the Challenges and Benefits of the Bug Bounty Ecosystem. In *32nd USENIX Security Symposium (USENIX Security '23)*, 2265–2282. Extended version available as arXiv:2301.04781 (2023).

Bae, J.; Kwon, S.; and Myeong, S. 2024. Enhancing Software Code Vulnerability Detection Using GPT-4o and Claude-3.5 Sonnet: A Study on Prompt Engineering Techniques. *Electronics*, 13(13).

Bugcrowd Researcher Success Team. 2025. How and When to Effectively Escalate a Submission. https://www.bugcrowd.com/blog/how-and-when-to-effectively-escalate-a-submission/. Bugcrowd Blog, January 18, 2025.

Daily NK. 2025. North Korea ramps up cyber offensive: New research center to focus on AI-powered hacking. *Daily NK*. Accessed: 2025-03-21.

Deng, G.; Liu, Y.; Mayoral-Vilches, V.; Liu, P.; Li, Y.; Xu, Y.; Zhang, T.; Liu, Y.; Pinzger, M.; and Rass, S. 2024. PentestGPT: An LLM-empowered Automatic Penetration Testing Tool. arXiv:2308.06782.

Fang, R.; Bindu, R.; Gupta, A.; and Kang, D. 2024. Llm agents can autonomously exploit one-day vulnerabilities. *arXiv preprint arXiv:2404.08144*, 13: 14.

Ichnowski, J.; Chen, K.; Dharmarajan, K.; Adebola, S.; Danielczuk, M.; Mayoral-Vilches, V.; Zhan, H.; Xu, D.; Ghassemi, R.; Kubiatowicz, J.; et al. 2023. Fogros 2: An adaptive and extensible platform for cloud and fog robotics using ros 2. In *Proceedings IEEE International Conference on Robotics and Automation*.

Kirschgens, L. A.; Ugarte, I. Z.; Uriarte, E. G.; Rosas, A. M.; and Vilches, V. M. 2018. Robot hazards: from safety to security. *arXiv preprint arXiv:1806.06681*.

Lera, F. J. R.; Santamarta, M. Á. G.; Costales, G. E.; Ayucar, U.; Gil-Uriarte, E.; Glera, A.; and Mayoral-Vilches, V. 2022. Threat modeling for robotic-based production plants. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 110–115. IEEE.

Maggi, F.; Vosseler, R.; Cheng, M.; Kuo, P.; Toyama, C.; Yen, T.; and Vilches, E. B. V. 2022. A Security Analysis of the Data Distribution Service (DDS) Protocol. *Trend Micro Research*.

Mayoral-Vilches, V. 2022. Robot cybersecurity, a review. *International Journal of Cyber Forensics and Advanced Threat Investigations*.

Mayoral-Vilches, V.; Deng, G.; Liu, Y.; Pinzger, M.; and Rass, S. 2023. ExploitFlow, cyber security exploitation routes for Game Theory and AI research in robotics. *arXiv e-prints*, arXiv–2308.

Mayoral-Vilches, V.; García-Maestro, N.; Towers, M.; and Gil-Uriarte, E. 2020. DevSecOps in Robotics. *arXiv preprint arXiv:2003.10402*.

Mayoral-Vilches, V.; Glera-Picón, A.; Ayúcar-Carbajo, U.; Rass, S.; Pinzger, M.; Maggi, F.; and Gil-Uriarte, E. 2021. Hacking planned obsolescense in robotics, towards security-oriented robot teardown. *Electronic Communications of the EASST*, 80.

Mayoral-Vilches, V.; Glera-Picón, A.; Ayucar-Carbajo, U.; Rass, S.; Pinzger, M.; Maggi, F.; and Gil-Uriarte, E. 2022a. Robot teardown, stripping industrial robots for good. *International Journal of Cyber Forensics and Advanced Threat Investigations*.

Mayoral-Vilches, V.; Juan, L. U. S.; Carbajo, U. A.; Campo, R.; de Cámara, X. S.; Urzelai, O.; García, N.; and Gil-Uriarte, E. 2019. Industrial robot ransomware: Akerbeltz. *arXiv preprint arXiv:1912.07714*.

Mayoral-Vilches, V.; Navarrete-Lozano, L. J.; Sanz-Gómez, M.; Espejo, L. S.; Crespo-Álvarez, M.; Oca-Gonzalez, F.; Balassone, F.; Glera-Picón, A.; Ayucar-Carbajo, U.; Ruiz-Alcalde, J. A.; et al. 2025. CAI: An Open, Bug Bounty-Ready Cybersecurity AI. *arXiv preprint arXiv:2504.06017*.

Mayoral-Vilches, V.; White, R.; Caiazza, G.; and Arguedas, M. 2022b. Sros2: Usable cyber security tools for ros 2. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11253–11259. IEEE.

Mayoral-Vilches, V. M.; Kirschgens, L. A.; Calvo, A. B.; Cordero, A. H.; Pisón, R. I.; Vilches, D. M.; Rosas, A. M.; Mendia, G. O.; Juan, L. U. S.; Ugarte, I. Z.; et al. 2018. Introducing the robot security framework (rsf), a standardized methodology to perform security assessments in robotics. *arXiv preprint arXiv:1806.04042*.

Microsoft. 2025. How AI is Transforming Cybersecurity: Tackling the Surge in Cyber Threats. Accessed: 2025-03-28.

OpenAI. 2025. o3-mini System Card. OpenAI Technical Report.

Rass, S.; König, S.; Wachter, J.; Mayoral-Vilches, V.; and Panaousis, E. 2023. Game-theoretic APT defense: An experimental study on robotics. *Computers & Security*, 132: 103328.

Shao, M.; Jancheska, S.; Udeshi, M.; Dolan-Gavitt, B.; Xi, H.; Milner, K.; Chen, B.; Yin, M.; Garg, S.; Krishnamurthy, P.; Khorrami, F.; Karri, R.; and Shafique, M. 2025. NYU CTF Bench: A Scalable Open-Source Benchmark Dataset for Evaluating LLMs in Offensive Security. arXiv:2406.05590.

Sridhar, K.; and Ng, M. 2021. Hacking for good: Leveraging HackerOne data to develop an economic model of bug bounties. *Journal of Cybersecurity*, 7(1): tyab007.

Staff, R. 2025. OpenAI removes users from China, North Korea over suspected malicious activities. *Reuters*. Accessed: 2025-03-20.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yen, T.-L.; Maggi, F.; Boasson, E.; Mayoral-Vilches, V.; Cheng, M.; Kuo, P.; and Toyama, C. 2021. The Data Distribution Service (DDS) Protocol is Critical Let's Use it Securely. *Blackhat EU*.