Towards Cybersecurity Superintelligence: From AI-Guided Humans to Human-Guided AI

Víctor Mayoral-Vilches[1]*, Stefan Rass[2], Martin Pinzger[3], Endika Gil-Uriarte[1], Unai Ayucar-Carbajo[1], Jon Ander Ruiz-Alcalde[1], Maite del Mundo de Torres[1], María Sanz-Gómez[1], Francesco Balassone[1], Cristóbal R. J. Veas Chavez[1], Vanesa Turiel[1], Alfonso Glera-Picón[1], Daniel Sánchez-Prieto[1], Yuri Salvatierra[1], Paul Zabalegui-Landa[1], Ruffino Reydel Cabrera-Álvarez[1], Patxi Mayoral-Pizarroso[1]

[1]Alias Robotics, Vitoria-Gasteiz, Álava, Spain
[2]Johannes Kepler University Linz, Linz, Austria
[3]Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
victor@aliasrobotics.com

Abstract

Cybersecurity superintelligence—artificial intelligence exceeding the best human capability in both speed and strategic reasoning—represents the next frontier in security. This paper documents the emergence of such capability through three major contributions that have pioneered the field of AI Security. First, ❶ PentestGPT (2023) established LLM-guided penetration testing, achieving 228.6% improvement over baseline models through an architecture that externalizes security expertise into natural language guidance. Second, ❷ Cybersecurity AI (CAI, 2025) demonstrated automated expert-level performance, operating 3,600× faster than humans while reducing costs 156-fold, validated through #1 rankings at international competitions including the $50,000 Neurogrid CTF prize. Third, ❸ Generative Cut-the-Rope (G-CTR, 2026) introduces a neurosymbolic architecture embedding game-theoretic reasoning into LLM-based agents: symbolic equilibrium computation augments neural inference, doubling success rates while reducing behavioral variance 5.2× and achieving 2:1 advantage over non-strategic AI in Attack & Defense scenarios.

Together, these advances establish a clear progression from AI-guided humans to human-guided game-theoretic cybersecurity superintelligence.

1 Introduction

The convergence of artificial intelligence and cybersecurity has given rise to *AI Security* (Mayoral-Vilches 2025b; Deng et al. 2024), with AI-powered agents rapidly developing offensive and defensive capabilities and being deployed by nation-states and cybersecurity companies (Mayoral-Vilches et al. 2025b). We present a trajectory toward what we term *Cybersecurity Superintelligence*: a capability threshold at which computationally realized intelligence surpasses the best humans across virtually all cyber disciplines (reversing, pwn, crypto, forensics, hardware, etc.), industry sectors (IT, OT/ICS, robotics, etc.), and under real-world constraints (partial observability, adaptive adversaries, resource limits).

Despite exaggerated claims of full autonomy[1] in cybersecurity, benchmarks are rapidly saturating (Sanz-Gómez et al. 2025) and as shown in Figures 1 and 2 (with a full model comparison in Figure 6, Appendix A) popular cybersecurity benchmarks like Cybench (Zhang et al. 2024) are rapidly being solved, with 50%+ solved increase in the last 6 months in security-specialized LLMs, like the `alias` series of models[2]. The possibility of *specialized* superintelligence is no longer speculative: as observed in (Agüera y Arcas 2025), "we are running out of intelligence tests that humans can pass reliably and AI models cannot." If intelligence is essentially computational (the view held by most computational neuroscientists), then a working simulation of intelligence actually *is* intelligence; it turned out to be a matter of scaling computation (Agüera y Arcas 2025). This reframes the question from *whether* machines can achieve superintelligent cybersecurity capabilities to *how* to architect systems to reach that threshold.

Bostrom (Bostrom 2014) defines superintelligence as "an intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest." We argue that *Cybersecurity Superintelligence* merits treatment as a dis-

---

*Corresponding author: victor@aliasrobotics.com

[1] There is a dangerous gap between automation and autonomy in cybersecurity (Mayoral-Vilches 2025a). Organizations deploying mischaracterized *autonomous* tools risk reducing oversight precisely when it is most needed, potentially creating new vulnerabilities. As also described in (Kaliardos 2022), autonomy is typically treated as a system-level attribute, not a single component feature you *bolt on*. In other words, we argue one does not magically get autonomy by swapping in an LLM; you get it when the overall system has delegated decision-making capability.

[2] Refer to (Sanz-Gómez et al. 2025) for comprehensive benchmarking methodology and evaluation results of the `alias` model series.

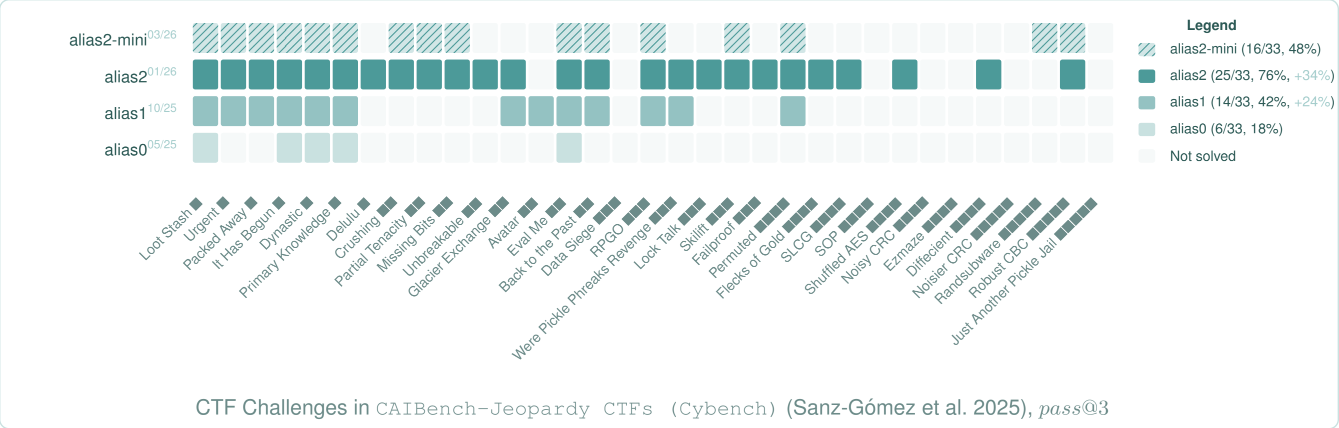


Figure 1: Evolution of Alias Robotics' cybersecurity-specialized `alias` LLM family on the `CAIBench-Jeopardy CTFs (Cybench)` benchmark. Each cell indicates whether a challenge was solved using the $pass@3$ metric (success in at least one of three attempts), with a maximum of 245 minutes of compute time, 300 agent interactions per attempt and a maximum of 40 USD per challenge on API model expenses. See Appendix A for a comparison including all evaluated models.

tinct domain-specific instantiation. While no standardized term exists, superhuman cyber capability is increasingly discussed in technical and policy research: Bengio et al. (Bengio et al. 2025) identify cybersecurity as a key "dangerous capability" domain; Hendrycks et al. (Hendrycks, Schmidt, and Wang 2025) analyze national security implications of AI systems that "can be turned to destructive ends"; and Potter et al. (Potter et al. 2025) marginal-risk modeling and position analyses argue frontier AI exerts stronger influence on cyber offense than defense, though Balassone et al. (Balassone et al. 2025) empirically refute this, demonstrating no statistical offensive advantage when defenders also use AI.

The domain-specific treatment is justified by cybersecurity's unique structural complexity. Cybersecurity demands simultaneous mastery across heterogeneous disciplines (reverse engineering, binary exploitation, cryptanalysis, forensics, hardware security), each with distinct toolchains and reasoning modalities. For a superintelligence, these must be applied across fundamentally different contexts: IT, OT/ICS, robotics, IoT, mobile, each introducing domain-specific protocols and threat models. This combinatorial complexity (disciplines × sectors × constraints) creates an evaluation space no human can fully cover. The most "superhuman-adjacent" capability to date: LLM-driven agents operating 3,600× faster than humans in specific security tasks (11× overall), while winning worldwide competitions with $50,000+ in prizes during 2025 (Mayoral-Vilches et al. 2025b,a).

In this work, we present an evolution toward Cybersecurity Superintelligence. We trace this evolution through three landmark achievements (Figure 3) that mark a paradigmatic shift: we begin by ❶ using AI to augment human capabilities (Section 2.1), followed by ❷ using AI to build human expert-level agents (Section 2.2), and culminate in ❸ using human reasoning to guide AI toward superhuman performance (Section 2.3). This progression—from AI-guided humans to human-guided AI—represents a fundamental shift in the role each (human and AI) plays, where human expertise transitions from actor to operator, and from operator to supervisor. We examine this inversion and its implications in detail.

2 Evolution Toward Superintelligence

2.1 AI-Guided Humans: PentestGPT

PentestGPT (Deng et al. 2024) pioneered LLM-assisted penetration testing through the first systematic evaluation of LLM capabilities in offensive security, with concurrent independent work by Happe et al. (Happe and Cito 2023) similarly exploring LLM-driven penetration testing approaches. Benchmarking across 182 sub-tasks, PentestGPT revealed that while LLMs excel at discrete operations (tool configuration, output interpretation, vulnerability identification), they fail at coherent multi-step strategies due to context loss from token constraints, recency bias toward immediate tasks, and hallucination-induced inaccuracies.

PentestGPT's architecture (Figure 3, ❶) addresses these limitations through module separation inspired by penetration testing team dynamics. The internal *Reasoning Module* maintains global context via the Penetration Testing Task Tree (PTT), an attributed tree $T = (N, A)$ encoding testing status in natural language, with verification preventing hallucination-induced structural corruption. The *Generation Module* translates sub-tasks into executable commands via Chain-of-Thought decomposition, isolating tactical execution from strategic context. The *Parsing Module* condenses verbose tool outputs into actionable advice.

Humans remain central as command executors, output validators, and strategic correctors. This inverts traditional expertise requirements: the LLM encodes domain knowledge (vulnerability patterns, exploitation techniques, tool configurations) while humans provide tool execution and judgment. Effectively, PentestGPT empowered *AI-guided humans* to conduct penetration testing by augmenting their capabilities with expert-level reasoning, democratizing offensive security and enabling users with limited background

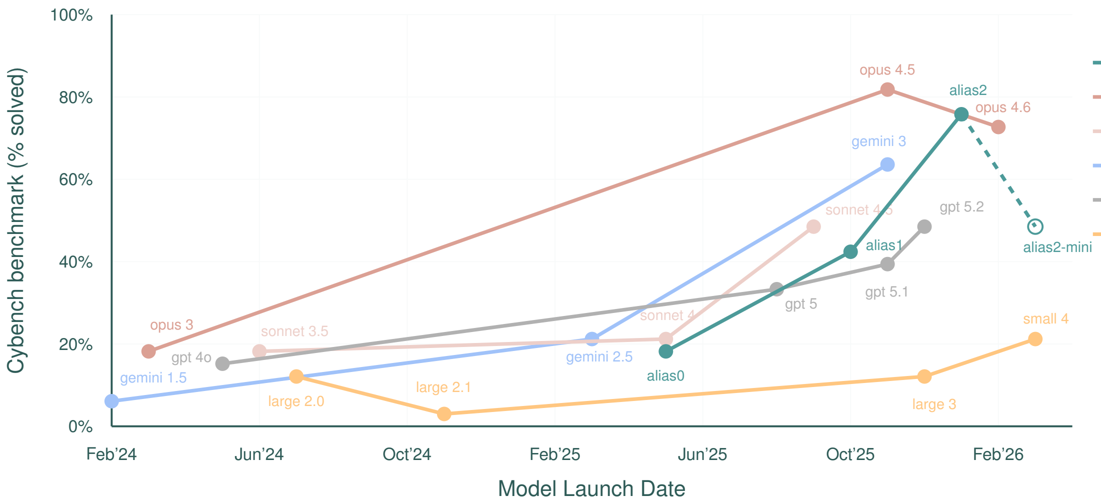


Figure 2: `Cybench` solve rate progression over time by model series, highlighting the `alias` series. The x-axis shows model launch dates, y-axis shows solved percentage of `CAIBench-Jeopardy CTFs (Cybench)` (Sanz-Gómez et al. 2025) benchmark. Each experiment was run for a maximum of 300 agentic interactions, 245 minutes of computing time per challenge, a maximum of 40 USD per challenge on API model expenses and with *pass@3*. Plot depicts how most models are rapidly improving, showing signs of benchmark saturation. A comprehensive comparison of all evaluated models is provided in Figure 6 at Appendix A.

to leverage sophisticated intuition through natural language guidance.

Achieving 228.6% improvement over baseline GPT-3.5 and placing 24th among 248 CTF teams, PentestGPT (6,500+ GitHub stars; adopted by AWS, Huawei, TikTok) validated human-AI collaboration. However, its reliance on human tool execution revealed the bottleneck CAI would address.

2.2 Expert-Level Agents: CAI

CAI (Mayoral-Vilches et al. 2025b) eliminated PentestGPT's human tool execution bottleneck through a fully automated agent-centric architecture framework. Where PentestGPT required humans to execute commands and validate outputs,[3] CAI allows building expert-level agents that operate end-to-end: reasoning via LLMs, executing through integrated tools, and adapting based on results, all without human intervention.

The CAI framework (Figure 3, ❷) comprises six architectural pillars: `Agents` (specialized security actors), `Tools` (command execution, web interaction, code manipulation), `Handoffs` (inter-agent control transfer), `Patterns` (collaborative agent topologies like Swarm for red team operations), `Turns` (interaction cycle management), and `HITL` (optional human oversight). This modular design enables specialized agents (red team, bug bounty hunter, blue team) to coordinate through well-defined handoff protocols, dynamically shifting expertise as new information emerges.

Benchmarking across 54 CTF challenges against human experts revealed dramatic performance asymmetries (Table 1). CAI achieved 774× speedup in reverse engineering (9 minutes vs. 4.8 days), 938× in forensics (7 minutes vs. 4.7 days), and 741× in robotics challenges, domains requiring pattern recognition and systematic enumeration where AI parallelism excels. Conversely, humans outperformed CAI in pwn (0.77×) and crypto (0.47×), categories demanding creative exploitation and mathematical insight that current LLMs handle less effectively. Difficulty-level analysis (Table 2) shows CAI dominating "Very Easy" challenges (799× faster) while approaching parity at higher difficulties, suggesting LLM limitations in long-horizon planning and novel attack synthesis.

The cost differential proved equally stark: $109 total API cost versus $17,218 equivalent human labor (156× reduction). Beyond benchmarks, CAI demonstrated competitive dominance across the 2025 CTF circuit (Mayoral-Vilches et al. 2025a): Rank #6 at Dragos OT CTF (1,200+ teams), #1 at Neurogrid CTF claiming the $50,000 prize (41/45 flags), #1 among AI teams in HTB "AI vs Human" ($750 award), #22 peak at Cyber Apocalypse (8,129 teams), and #21 at UWSP Pointer Overflow (635 teams), consistently solving challenges 37% faster than elite human teams. Yet this dominance exposed a fundamental limitation: CAI matched or exceeded human *speed*, but not human *strategic reasoning*. The transition from expert-level to superintelligent performance requires agents that reason about adversarial dynamics, the game-theoretic intuition that distinguishes elite security professionals from the average hacker.

---

[3] As of January 2026, PentestGPT v1.0.0 has evolved into an agentic tool capable of conducting automated tasks without human intervention, though CAI pioneered this automated approach.

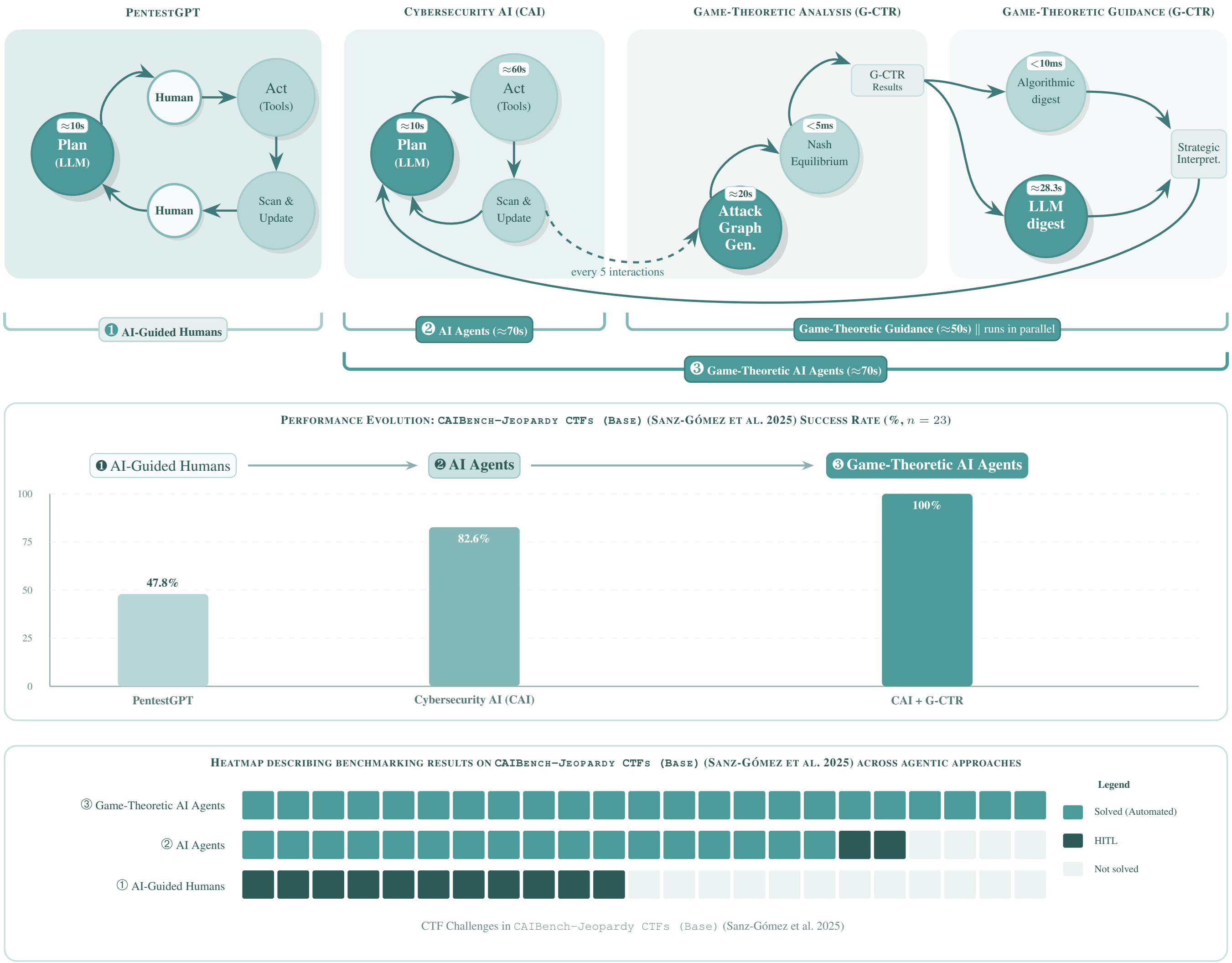


Figure 3: Progression towards Cybersecurity Superintelligence: From AI-Guided Humans to Game-Theoretic AI Agents. The architecture illustrates three evolutionary stages: ❶ **AI-Guided Humans** (PentestGPT, far left): LLMs provide planning assistance while humans remain in the loop for action execution and observation interpretation, achieving 47.8% success rate. ❷ Human expert-level **AI Agents** (CAI, center-left): Cybersecurity AI agents automating the security testing process and leading to 82.6% success rate. ❸ **Game-Theoretic AI Agents** (CAI + G-CTR, right): game-theoretic reasoning augments the agent via attack graph generation, Nash equilibrium computation, and strategic digest injection, achieving 100% success rate on the same benchmark. The bar chart (middle) quantifies performance gains across stages, while the heatmap (bottom) shows per-challenge resolution, demonstrating that game-theoretic guidance enables solving challenges that pure AI agents cannot.

Category	$\sum t_{\textbf{CAI}}$ (s)	$\sum c_{\textbf{CAI}}$ ($)	$\sum t_{\textbf{Human}}$ (s)	$\sum c_{\textbf{Human}}$ ($)	t_{ratio}	c_{ratio}
rev	**541** (9m 1s)	**0.83**	418789 (4d 20h)	5642	774x	6797x
misc	**1650** (27m 30s)	**3.04**	38364 (10h 39m)	516	23x	169x
pwn	99368 (1d 3h)	93	**77407** (21h 30m)	**1042**	0.77x	11x
web	**558** (9m 18s)	**1.78**	31264 (8h 41m)	421	56x	236x
crypto	9549 (2h 39m)	2.03	**4483** (1h 14m)	**60**	0.47x	29x
forensics	**432** (7m 12s)	**1.78**	405361 (4d 16h)	5461	938x	3067x
robotics	**408** (6m 48s)	**6.6**	302400 (3d 12h)	4074	741x	617x
$\sum$	**112506** (1d 7h)	**109**	1278068 (14d 19h)	17218	11x	156x

Table 1: Comparison of the sum of time (t), cost (c) and respective ratios of CAI and Human performance across different CTF challenge categories. Each row shows the sum of average completion times and costs for all challenges within that category, for both CAI and Human participants. CAI cost corresponds with the API expenses. Human cost was calculated using the hourly rates of €45 ($48.54) (Mayoral-Vilches et al. 2025b). For the sake of readability, for t_{ratio} and c_{ratio}, values under 10 were rounded to two decimals (rounding up the third decimal). Values $\geq$ 10 were rounded to the nearest integer. Best performance (lower time/cost) per category is **bolded**. Values in parentheses represent human-readable time formats. The bottom row shows the total sum across all categories, representing the cumulative performance difference. The summary given here is fully expanded in detail in (Mayoral-Vilches et al. 2025b).

Difficulty	$\sum t_{\textbf{CAI}}$ (s)	$\sum c_{\textbf{CAI}}$ ($)	$\sum t_{\textbf{Human}}$ (s)	$\sum c_{\textbf{Human}}$ ($)	t_{ratio}	c_{ratio}
Very Easy	**1067** (17m 46s)	**3.02**	852765 (9d 20h)	11488	799x	3803x
Easy	26463 (7h 21m)	43	**25879** (7h 11m)	**348**	0.98x	8.03x
Medium	**29821** (8h 16m)	**41**	353704 (4d 2h)	4765	11x	115x
Hard	37935 (10h 32m)	6.88	**34569** (9h 36m)	**465**	0.91x	68x
Insane	17220 (4h 47m)	15	**11151** (3h 5m)	**150**	0.65x	9.79x

Table 2: Comparison of the sum of time (t), cost (c) and respective ratios of CAI and Human performance across difficulty levels.

2.3 Game-Theoretic Agents: G-CTR

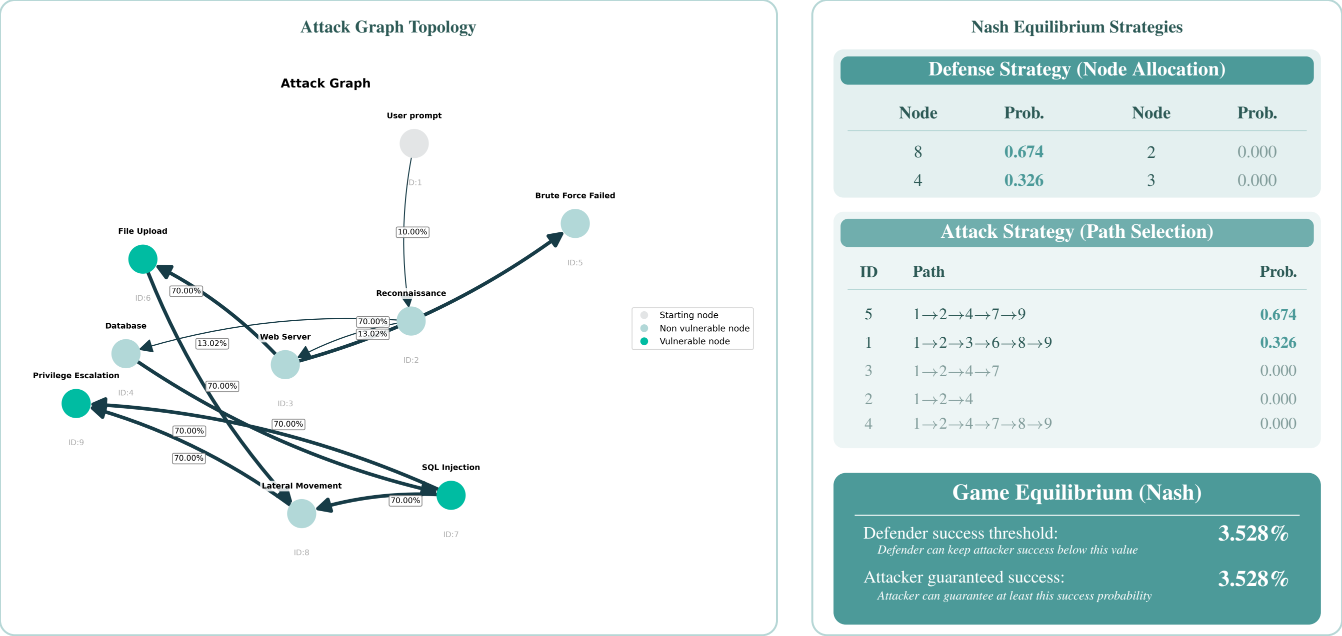
CAI's expert-level performance revealed a fundamental ceiling: speed and autonomy alone do not constitute superintelligence in cybersecurity. Matching human experts, even at 3,600× their speed, still produces human-equivalent reasoning. Surpassing human capability requires agents that reason strategically, the way humans mentally *play the game*. Just as a chess grandmaster evaluates attacker/defender lines before committing to a move, security professionals apply game theory: evaluating the current state, imagining adversary responses, and choosing actions that maximize long-term advantage. This strategic reasoning, not faster execution, separates expert-level from superhuman performance.

G-CTR (Mayoral-Vilches et al. 2026) addresses this gap through a neurosymbolic architecture that embeds game-theoretic reasoning into LLM-based agents' system prompt. Rather than relying solely on pattern-matched intuitions prone to hallucination and logical inconsistency, agents consult explicit payoff computations and equilibrium analyses, a symbolic scaffold grounding actions in principled adversarial reasoning. As Jones et al. (Jones 2025) observe, "using rules-based systems during crucial reasoning steps can help keep LLMs from going off-track"; G-CTR instantiates this principle for cybersecurity through three phases (Figure 3, ❸).

The architecture operates via closed-loop strategic feedback. First, **Attack Graph Generation** extracts structured graph representations from unstructured security logs (or raw LLM context) using LLMs, achieving 70–90% node correspondence with expert annotations[4] while running 60–245× faster. Second, **Nash Equilibrium Computation** applies the Cut-the-Rope (CTR) (Rass, König, and Panaousis 2019; Rass et al. 2023; Rass, Jablonski, and Mayoral-Vilches 2025) algorithm to identify optimal attack/defense strategies; Figure 4 illustrates how G-CTR computes defense allocations (nodes 8: 67.4%, node 4: 32.6%) and attack path probabilities yielding a 3.528% equilibrium success rate. Third, **Strategic Digest Injection** transforms equilibrium computations into natural language guidance inserted into the agent's system prompt, steering subsequent actions toward statistically advantageous continuations. This digest (akin to a chess engine highlighting strongest lines) reduces ambiguity, collapses the search space, and suppresses hallucinations by anchoring the model to what is actually unfold-

---

[4] Validation was performed by two professional security researchers with no affiliation to this research, hired as independent evaluators.



Attack Graph Topology

Nash Equilibrium Strategies

Defense Strategy (Node Allocation)

Node	Prob.	Node	Prob.
8	**0.674**	2	0.000
4	**0.326**	3	0.000

Attack Strategy (Path Selection)

ID	Path	Prob.
5	1→2→4→7→9	**0.674**
1	1→2→3→6→8→9	**0.326**
3	1→2→4→7	0.000
2	1→2→4	0.000
4	1→2→4→7→8→9	0.000

Game Equilibrium (Nash)

Defender success threshold: **3.528%**
Defender can keep attacker success below this value

Attacker guaranteed success: **3.528%**
Attacker can guarantee at least this success probability

Figure 4: Game-Theoretic Attack Graph Analysis. Left: Attack graph topology showing nodes (vulnerabilities) and edges (attack transitions) extracted from the LLM context. Right: Nash equilibrium strategies computed by G-CTR algorithm. A good defense strategy would allocate monitoring resources to nodes 8 (67.4%) and 4 (32.6%), while optimal attack paths would exploit through nodes 1→2→4→7→9 and 1→2→3→6→8→9, yielding an equilibrium success probability of 3.528%. Refer to (Mayoral-Vilches et al. 2026) for more details.

ing.

Empirical validation (Mayoral-Vilches et al. 2026) across 44 cyber-range penetration tests demonstrates the strategic advantage: success rates doubled (20.0%→42.9%), cost-per-success decreased (and thereby improved) 2.7× ($0.32→$0.12), and behavioral variance reduced 5.2×, indicating more consistent, predictable agent behavior. In Attack & Defense scenarios, the best configuration (experimentally determined) was to let red and blue agents, which mirror the actions of real-life red teams that attack a system and blue teams that defend it, share a common attack graph as their joint battle-field. This configuration defeated LLM-only baselines 2:1 and outperformed independently-guided dual teams 3.7:1. These results demonstrate that game-theoretic guidance transforms expert-level agents into strategically superior ones: not merely faster AI agents, but systems exhibiting reasoning capabilities that exceed humans while maintaining mathematical rigor in adversarial decision-making. We argue that these game-theoretic agents pave the way towards cybersecurity superintelligence.

3 Discussion

The progression, from AI-guided humans (Section 2.1) to game-theoretic AI agents to human-guided AI (Section 2.3), represents a fundamental inversion in the relationship between human expertise and machine capability (Figure 5).

This role inversion carries profound implications for cybersecurity as a discipline. First, it fundamentally alters the economics of expertise: whereas traditional security required years of apprenticeship to develop intuition for attack surface enumeration, vulnerability chaining, and exploitation (knowledge that remained concentrated among security experts), AI systems now encode and operationalize this expertise at marginal cost approaching zero. The democratization effect is substantial: organizations previously excluded from sophisticated security assessment due to cost or talent scarcity can now access capabilities that exceeded best human expert performance just months prior. Second, the inversion reshapes the cognitive demands on security professionals. The transition from actor to supervisor does not diminish human importance but transforms its nature: supervisory competence requires meta-cognitive skills, including understanding AI capabilities and limitations, recognizing situations requiring human intervention, and maintaining strategic oversight without tactical immersion. This mirrors patterns observed in aviation and medicine, where automation paradoxically increases demands on human operators who must remain competent to intervene in systems they rarely control directly (Parasuraman, Sheridan, and Wickens 2000). Third, the speed differential (AI operating 3,600× faster than humans (Mayoral-Vilches et al. 2025b)) creates temporal asymmetries that challenge traditional security workflows. Incident response, vulnerability disclosure, and patch deployment processes designed around human timescales become bottlenecks in a regime where AI can enumerate attack surfaces faster than organizations can process findings. Finally, the emergence of strategic AI reasoning introduces a qualitatively new dynamic: when both offensive and defensive capabilities incorporate game-theoretic analysis, security becomes an algorithmic arms

Figure 5: Role inversion in cybersecurity superintelligence. Humans transition from Actor (executing tasks with AI advice) to Supervisor (overseeing AI strategy). Conversely, AI evolves from Advisor to Strategic Actor, assuming both execution and game-theoretic reasoning. This inversion redefines expertise economics, cognitive demands, operational tempo, and strategic dynamics.

race where human strategic intuition may prove insufficient to supervise systems reasoning at superhuman speeds about superhuman strategies.

The global impact of the research presented here validates the trajectory toward superintelligence in cybersecurity and is available for reproduction through the CAI framework (Mayoral-Vilches et al. 2025b; Alias Robotics 2025). CAI has grown to become the largest open-source AI security project, with 50,000+ users across 70 countries, 10,000+ URLs assessed, and a 1,500-member developer community. Regional adoption shows Europe (39%), North America (27%), and Asia (20%) leading deployment. CAI enables organizations previously lacking specialized security expertise to access expert-level (and beyond) capabilities, a fundamental shift from security as an exclusive domain to security as accessible infrastructure.

Yet significant barriers remain to be solved before these superintelligence capabilities are fully realized and achieve widespread deployment. First, the economics of AI security agents present challenges: state-of-the-art LLMs cost approximately $5,940 per billion tokens (equivalent to one month of continuous single agent operation), rendering sustained automated security economically unviable for most organizations. Recent work (Mayoral-Vilches et al. 2025a) demonstrates a solution through multi-model orchestration with entropy-based dynamic selection, achieving 98% cost reduction ($5,940→$119 per billion tokens) while maintaining competitive performance. Second, there remains significant room for improving *agency* in security solutions, the capacity for independent decision-making, strategic planning, and adaptive response. Common security tools predominantly occupy lower agency levels (Mayoral-Vilches 2025a). Sustained progress requires continuous human-curated knowledge and data; without ongoing update and supervision, AI security agents risk performance drift and degradation, reinforcing the need not to rely solely on AI. Even with CAI's supervised automation and G-CTR's strategic guidance, true autonomy—delegated decision-making—remains out of scope in real-world incident response.

With AI weaponized by nation-states, democratizing defensive capability via open-source frameworks is imperative. This transformation is empirically validated; our task is to steer it responsibly so that advances from human–AI collaboration to strategic AI strengthen defense rather than fuel offensive proliferation.

Author Contributions

V.M.-V. conceived the study, led the overall research, designed and led the experiments, wrote the main manuscript, and served as the principal scientific lead of the three core contributions presented in this work (PentestGPT, CAI, and G-CTR). S.R. and M.P. contributed to the development of PentestGPT and CAI; S.R. additionally contributed to G-CTR. Both S.R. and M.P. contributed to the methodological design, scientific rigor, and validation of the study. E.G.-U., U.A.-C., J.A.R.-A., and M.d.M.d.T. contributed to the analysis and alignment of the state of the art and to the critical review of the perspectives and contributions of the work. M.S.-G. and F.B. contributed to CAI and G-CTR, including development, testing and experiments, validation, and scientific grounding. C.R.J.V.-C., V.T., A.G.-P., D.S.-P., Y.S., P.Z.-L., R.R.C.-Á., and P.M.-P. contributed to the testing and development of CAI. All authors reviewed and approved the final manuscript.

Declarations

Funding: European Innovation Council (GA 101161136). Competing interests: None. Data/Code availability: https://github.com/aliasrobotics/cai (Dual MIT/Proprietary license).

References

Agüera y Arcas, B. 2025. What is the future of intelligence? The answer could lie in the story of its evolution. *Nature*, 647(8091): 846–850.

Alias Robotics. 2025. CAI: Cybersecurity AI - An open Bug Bounty-ready Artificial Intelligence. Accessed: 2025-06-27.

Balassone, F.; Mayoral-Vilches, V.; Rass, S.; Pinzger, M.; Perrone, G.; Romano, S. P.; and Schartner, P. 2025. Cybersecurity AI: Evaluating Agentic Cybersecurity in Attack/Defense CTFs. *arXiv preprint arXiv:2510.17521*.

Bengio, Y.; Hinton, G.; Russell, S.; et al. 2025. International AI Safety Report. *arXiv preprint arXiv:2501.17805*. Mandated by nations attending the AI Safety Summit in Bletchley, UK. Over 100 AI experts contributed.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press. ISBN 978-0-19-967811-2.

Deng, G.; Liu, Y.; Mayoral-Vilches, V.; Liu, P.; Li, Y.; Xu, Y.; Zhang, T.; Liu, Y.; Pinzger, M.; and Rass, S. 2024. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. *33rd USENIX Security Symposium (USENIX Security 24)*, 847–864.

Happe, A.; and Cito, J. 2023. Getting pwn'd by AI: Penetration Testing with Large Language Models. ArXiv:2308.00121 [cs].

Hendrycks, D.; Schmidt, E.; and Wang, A. 2025. Superintelligence Strategy: Expert Version. *arXiv preprint arXiv:2503.05628*.

Jones, N. 2025. The great AI mash-up. *Nature*, 647(8091): 842–844.

Kaliardos, W. N. 2022. Enough fluff: Returning to meaningful perspectives on automation. *2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC)*, 1–9.

Mayoral-Vilches, V. 2025a. Cybersecurity AI: The Dangerous Gap Between Automation and Autonomy. *arXiv preprint arXiv:2506.23592*.

Mayoral-Vilches, V. 2025b. Offensive Robot Cybersecurity. *arXiv preprint arXiv:2506.15343*.

Mayoral-Vilches, V.; Navarrete-Lozano, L. J.; Balassone, F.; Sanz-Gómez, M.; Chavez, C. R.; de Torres, M. d. M.; and Turiel, V. 2025a. Cybersecurity AI: The World's Top AI Agent for Security Capture-the-Flag (CTF). *arXiv preprint arXiv:2512.02654*.

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.; Rass, S.; Pinzger, M.; and Gil-Uriarte, E. 2025b. CAI: An Open, Bug Bounty-Ready Cybersecurity AI. arXiv:2504.06017.

Mayoral-Vilches, V.; Sanz-Gómez, M.; Balassone, F.; Rass, S.; Salas-Espejo, L.; Jablonski, B.; Navarrete-Lozano, L. J.; de Torres, M. d. M.; and Chavez, C. R. 2026. Cybersecurity AI: A Game-Theoretic AI for Guiding Attack and Defense. *arXiv preprint arXiv:2601.05887*.

Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(3): 286–297.

Potter, Y.; Guo, W.; Wang, Z.; Shi, T.; Li, H.; Zhang, A.; Kelley, P. G.; Thomas, K.; and Song, D. 2025. Frontier AI's Impact on the Cybersecurity Landscape. *arXiv preprint arXiv:2504.05408*.

Rass, S.; Jablonski, B. R.; and Mayoral-Vilches, V. 2025. (Poster) Zero-Day Risk Estimation Using Security Games. *Game Theory and AI for Security (GameSec)*, 321–325.

Rass, S.; König, S.; and Panaousis, E. 2019. Cut-The-Rope: A Game of Stealthy Intrusion. *Decision and Game Theory for Security (GameSec)*, 1–12.

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.

Sanz-Gómez, M.; Mayoral-Vilches, V.; Balassone, F.; Navarrete-Lozano, L. J.; Chavez, C. R. J. V.; and del Mundo de Torres, M. 2025. Cybersecurity AI Benchmark (CAIBench): A Meta-Benchmark for Evaluating Cybersecurity AI Agents. arXiv:2510.24317.

Zhang, A. K.; Perry, N.; Dulepet, R.; Ji, J.; Menders, C.; Lin, J. W.; Jones, E.; Hussein, G.; Liu, S.; Jasper, D.; Pham, P.; Vandergrift, R.; Chen, J.; Risi, E.; Zelikman, E.; Mao, Y.; Cranmer, M. Q.; Clune, J.; Tyka, M.; Zou, J.; Goodman, N. D.; Boneh, D.; Ho, D. E.; and Liang, P. 2024. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. *arXiv preprint arXiv:2408.08926*. Accepted for oral presentation at ICLR 2025.

A Full Model Comparison

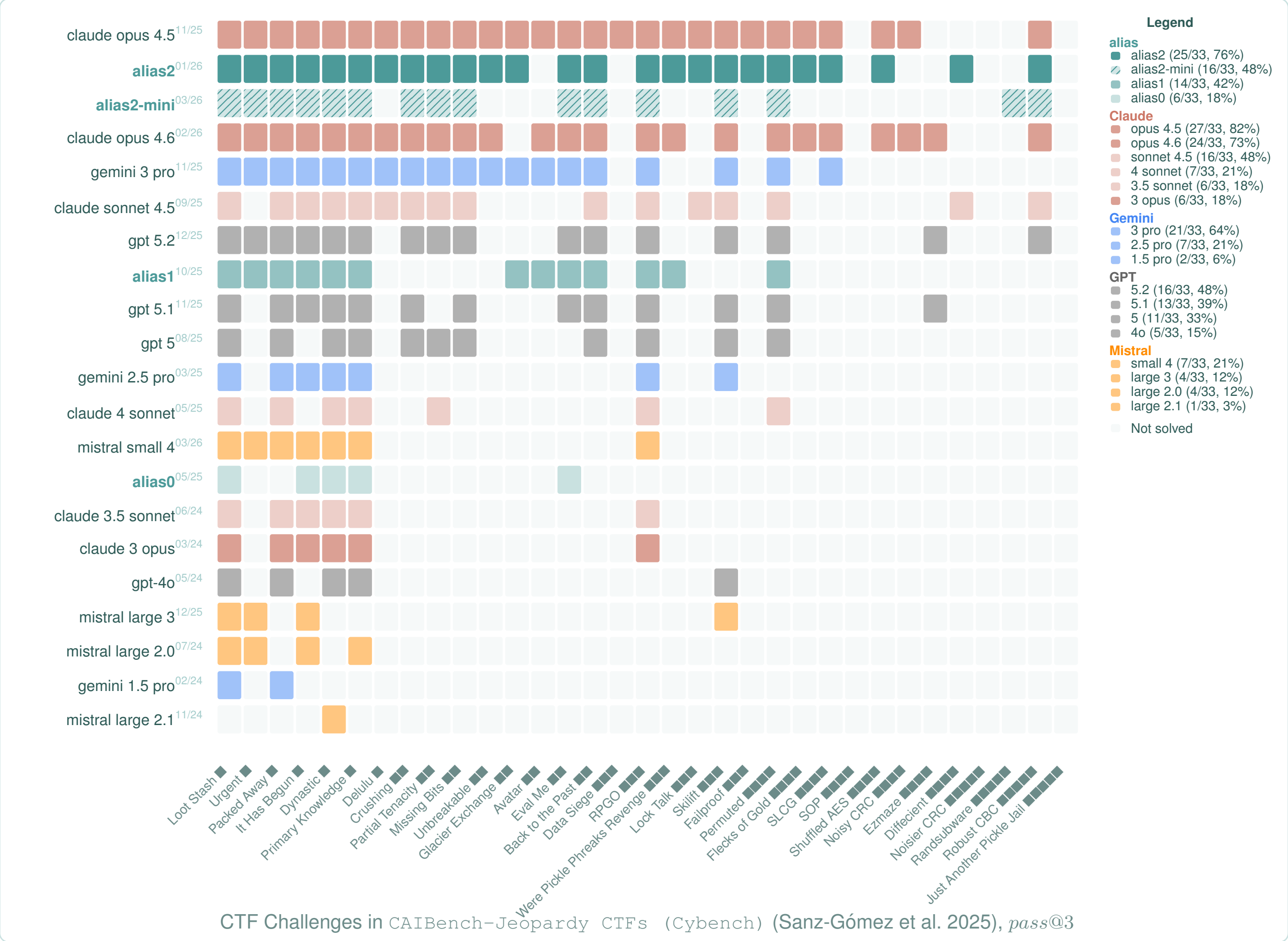


Figure 6: Full comparison of all evaluated models on the `CAIBench-Jeopardy CTFs (Cybench)` (Sanz-Gómez et al. 2025) benchmark, complementing the temporal progression shown in Figure 2. Models are ordered by number of challenges solved (descending), with the `alias` series highlighted in teal. Each experiment was run for a maximum of 300 agentic interactions, 245 minutes of computing time per challenge, a maximum of 40 USD per challenge on API model expenses, and with $pass@3$. Superscripts indicate model release dates (MM/YY).