Cybersecurity AI in OT: Insights from an AI Top-10 Ranker in the Dragos OT CTF 2025

Víctor Mayoral-Vilches[1], **Luis Javier Navarrete-Lozano**[1], **Francesco Balassone**[1], **María Sanz-Gómez**[1], **Cristóbal R. J. Veas Chavez**[1,2], **Maite del Mundo de Torres**[1] and **Endika Gil-Uriarte**[1]

[1] **Alias Robotics**, Vitoria-Gasteiz, Álava, Spain, ✉ research@aliasrobotics.com
[2] Alpen-Adria-Universität Klagenfurt.
⌸ https://github.com/aliasrobotics/cai , 🎮 https://discord.gg/fnUFcTaQAC

Abstract

Operational Technology (OT) cybersecurity increasingly relies on rapid response across malware analysis, network forensics, and reverse engineering disciplines. We examine the performance of Cybersecurity AI (CAI), powered by the `alias1` model, during the Dragos OT CTF 2025—a 48-hour industrial control system (ICS) competition with more than 1,000 teams. Using CAI telemetry and official leaderboard data, we quantify CAI's trajectory relative to the leading human-operated teams. CAI reached Rank 1 between competition hours 7.0 and 8.0, crossed 10,000 points at 5.42 hours (1,846 pts/h), and completed 32 of the competition's 34 challenges before automated operations were paused at hour 24 with a final score of 18,900 points (6th place). The top-3 human teams solved 33 of 34 challenges, collectively leaving only the 600-point "Kiddy Tags – 1" unsolved; they were also the only teams to clear the 1,000-point "Moot Force" binary. The top-5 human teams averaged 1,347 pts/h to the same milestone, marking a 37% velocity advantage for CAI. We analyse time-resolved scoring, category coverage, and solve cadence. The evidence indicates that a mission-configured AI agent can match or exceed expert human crews in early-phase OT incident response while remaining subject to practical limits in sustained, multi-day operations.

1 Introduction

The convergence of artificial intelligence and cybersecurity has reached an important milestone. While early work such as PentestGPT [1, 2] demonstrated that Large Language Models (LLMs) could assist human security practitioners, recent developments have shown that AI agents can operate mostly autonomously at near- and super-human levels [3]. However, most evaluations of cybersecurity AI have focused on generic IT-related penetration testing scenarios or knowledge-based benchmarks [4]. The question of how these AI systems perform in specialized Operational Technology (OT) contexts—where robots, industrial control systems, SCADA networks, and critical infrastructure are at stake—remains largely unexplored.
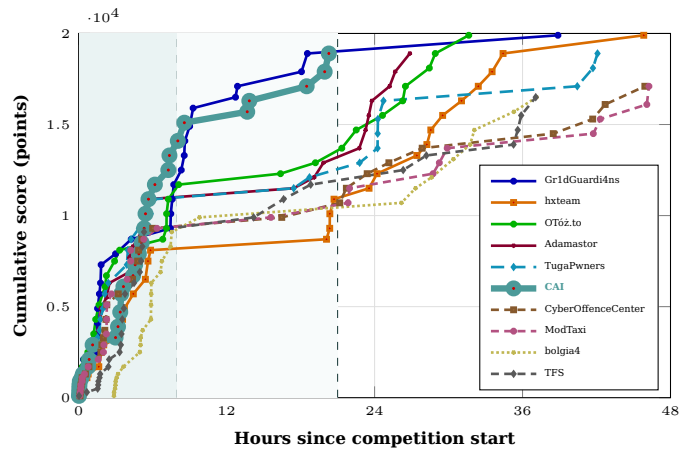

Figure 1: Top-10 trajectories across the 48-hour Dragos OT CTF 2025. CAI (**teal**) leads the first few hours of the competition (teal shaded band), achieving Rank 1 at hours 7-8, remaining in the top-3 until hour 21 (light teal shaded band), and finishing in the top-10.

This technical report presents a step towards addressing that gap by documenting CAI's performance in the Dragos OT CTF (Capture the Flag) 2025, a specialized competition focused on OT cybersecurity. The Dragos CTF attracts security practitioners, researchers, and teams from around the world who compete to solve industrial cybersecurity challenges involving malware analysis, network packet forensics, reverse engineering, and threat intelligence—all tailored to industrial control system environments. Dragos CTF is one of the largest OT-focused competitions in the world, and despite not being purposely built for AI evaluation, it provides a unique testbed for evaluating AI capabilities alongisde elite human teams in a domain where the stakes are considerably higher and skills significantly different than traditional IT security.

1.1 Motivation and Context

CAI moved beyond synthetic pilots during 2024 and early 2025, when we started deploying the framework against real industrial assets and observed autonomous end-to-end cybersecurity work within minutes. During the Ecoforest heat pump engagement, CAI enumerated the vendor's perimeter, recovered exposed `.htpasswd` files, cracked legacy DES hashes, and demonstrated full capabilities for remote manipulation of heating parameters without human intervention [5]. These results confirmed that OT-grade automation can already surface and weaponize critical flaws significantly faster than traditional assessment teams.

However, the same responsible-disclosure exercise revealed a sobering pattern previously observed in other vendors across robotics and wider OT technology domains [6]: the vendors don't react to the findings in a timely manner. Ecoforest, the vendor behind the ICS technologies found vulnerable was engaged in June 2025. The same information was conveyed to spanish national authorities via INCIBE [7], the Spanish national Cyber Security Institute and root CNA, also in June. At the time of writing, CAI reported—exposed services, default credentials, or unpatched internet-facing gateways remain vulnerable for more than 120 days after the first notification, even when remediation required only configuration changes rather than downtime. Persistent neglect of such basics underscores how vulnerable critical OT estates remain despite repeated warnings.

The combination of demonstrably autonomous[1] exploitation capabilities and lagging remediation among asset owners heightens the urgency for scalable OT security practices. Industrial control environments underpin heating, power distribution, manufacturing, and logistics, so latent exposures translate directly into physical risk (safety

Figure 2: Snapshot from CAI's semi-autonomous Ecoforest heat pump assessment: the agent recovers exposed `.htpasswd` credentials, cracks DES hashes, and prepares remote manipulation of heating parameters without human intervention. Full case study available at https://aliasrobotics.com/case-study-ecoforest.php.

hazards) and economic disruption. This reality drove us to keep investing in CAI's planning, tooling, and safety controls and, as part of that effort, to test the system against elite human operators in global OT challenges such as the Dragos OT CTF 2025. Our motivation is clear: to empower defenders to mitigate at the speed of attackers, if not faster. As recently discussed in AI Security Attack & Defense CTFs research [9], the current state of the art in AI-driven cybersecurity can empower defenders at similar
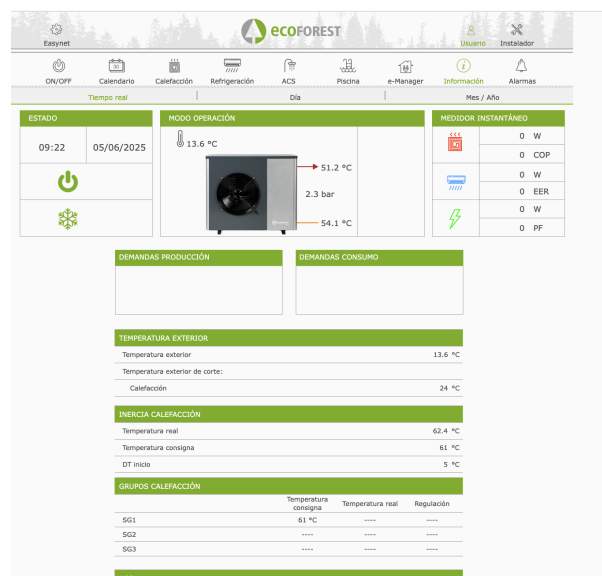
---

[1]Refer to [8] for the differences between autonomous and automated.

rates to attackers. Even with this motivation, structural headwinds continue to hamper defenders [6]:

- **Domain Complexity**: OT systems involve specialized protocols (e.g. Modbus, DNP3, IEC 61850), proprietary hardware, and unique operational constraints that differ fundamentally from IT systems.

- **Speed Requirements**: Industrial incidents can escalate quickly, requiring rapid threat analysis and response times that may exceed human capacity.

- **Expertise Scarcity**: OT cybersecurity talent is scarce, with relatively few vendors and highly demanded professionals possessing both deep industrial systems knowledge and advanced cyber threat analysis skills.

- **Scale Challenges**: Global critical OT infrastructure encompasses millions of devices, making manual security assessment and continuous monitoring highly challenging.

These forces shaped our research focus: quantify how far semi-autonomous agents can mitigate the speed, scale, and expertise gaps before remediation finally catches up. Figure 1 provides a high-level view of how CAI's early sprint translated into a day-one lead that we dissect in Section 3, while the remainder of this report examines the operational behaviour behind that curve Cybersecurity AI agents like the open source CAI [10] therefore represent a potential solution to these challenges, offering autonomous, scalable, and rapid security operations. The central question—tested throughout the Dragos OT CTF campaign—is how reliably those capabilities translate to contested, real-world OT scenarios.

1.2 Related Work

The application of AI to cybersecurity has evolved from assisted tools to increasingly autonomous systems. PentestGPT [1, 2] pioneered LLM-empowered penetration testing, achieving 228.6% improvement over GPT-3.5 baseline, though requiring significant human action. Subsequent work introduced AutoPenBench [11] with 33 challenges and multi-agent collaborative frameworks [12, 13], though evaluation revealed that even advanced models (Llama 3.1, GPT-4o) fall short of end-to-end autonomous penetration testing. Benchmarks have emerged to evaluate AI capabilities systematically: Cybench [14] established 40 professional-level CTF tasks from 2022-2024 competitions with objective difficulty ratings; EnIGMA [15] achieved state-of-the-art on 390 CTF challenges at ICML 2025, solving $3\times$ more than prior agents; CyberGym [16] addresses benchmark limitations by incorporating real-world complexities. CAIBench [4] provides meta-benchmark integration across 10,000+ instances, revealing 70% success on knowledge-based tasks but only 20-40% on multi-step adversarial scenarios.
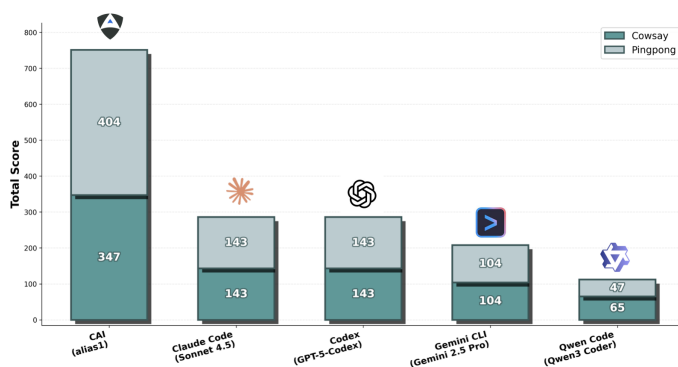

Figure 3: Benchmarking results of AI-vs-AI agents in attack/defense CTFs scenarios that simulate real-world security operations. CAI using `alias1` with the `red_teamer` and `blue_teamer` agents achieved state-of-the-art performance with a 2.6× speedup over the second best agent. Refer to [4] for more details.

AI systems have been evaluated in competitive contexts with mixed results. Hack The Box's "AI vs Human" competition [17] showed the best AI team (CAI) placing 20th out of 403 teams with 95% challenge completion (19/20 challenges). DARPA's AI Cyber Challenge [18] demonstrated autonomous infrastructure security capabilities with 7 finalists releasing open-source systems for vulnerability discovery and patching.

Originated from a PhD thesis [2], the CAI (Cybersecurity AI) framework

[3, 10] represents a progression toward autonomous cybersecurity operations. Prior CAI research established foundational capabilities: the distinction between automation and autonomy in cybersecurity AI [8], frameworks for evaluating AI fluency in security tasks [19], defensive techniques against prompt injection attacks on security AI systems [20], and evaluation in attack/defense CTF scenarios [9]. Additional work explored CAI applications to robotics security [21, 22]. However, prior evaluations focused primarily on IT security contexts or controlled benchmarks rather than real-world operational technology (OT) competitions.

This work differs from prior evaluations in four key aspects: (1) **OT/ICS specialization**—evaluation on operational technology with industrial protocols (Modbus, DNP3) and SCADA systems rather than generic IT security; (2) **competitive scale**—1,000+ teams in a live 48-hour competition rather than isolated benchmarks; (3) **sustained automated operation**—24 hours of CAI's problem-solving with minimal Human-In-The-Loop intervention; (4) **comparative performance**—documented achievement of Rank 1 position against elite human teams during the first 21 hours of the competition, to later rank in the top-10, providing empirical data on AI capabilities in time-critical, adversarial OT security scenarios.

1.3 Contributions

This technical report documents CAI's performance in a specialized OT security competition and makes the following contributions:

1. **OT Security Competition Results**: We document CAI's performance in the Dragos OT CTF 2025, a 48-hour OT-focused competition with 1,000+ teams. CAI achieved Rank 1 at hours 7-8 and finally finished the complete competition the top-10, solving 32 of 34 available challenges; only the top-three human teams cleared 33/34, leaving only the 600-point "Kiddy Tags – 1" unsolved while exclusively solving the 1,000-point "Moot Force" binary.

2. **Temporal Performance Analysis**: We present quantitative analysis of CAI's score progression based on real competition data. CAI achieved #1 velocity ranking at 1,846 pts/h to 10K points, 37.1% faster than the top-5 human team average (1,347 pts/h) demonstrating strong speed at performance. Analysis covers early-phase velocity (hours 0-7), milestone timing (1K at 0.24h, 5K at 3.45h, 10K at 5.42h), and the 24-hour automated operation window with minimal Human-In-The-Loop intervention.

The remainder of this report is organized as follows: Section 2 provides background on CAI, the `alias1` model, and the Dragos OT CTF 2025 competition structure. Section 3 presents our experimental methodology and evaluation approach, as well as detailed results including temporal analysis, comparative studies, and specific challenge examples. Section 4 discusses implications for OT cybersecurity and limitations of the current study. Finally, Section 5 concludes with reflections on the future of AI-driven industrial cyberdefense.

2 Background

2.1 CAI Framework and `alias1` Model

CAI (Cybersecurity AI) is an open-source framework for AI cybersecurity operations [3, 10, 8, 9]. Unlike AI-assisted tools [1, 12, 13], CAI operates fully automated with optional Human-In-The-Loop (HITL) supervision and through a multi-agent architecture powered by `alias1`, a specialized state of the art[4]

LLM optimized for security tasks including artifact analysis, hypothesis generation, strategic planning, and adaptive learning from feedback.

CAI was deployed with `alias1` (temperature 0.7, 200K token context), using the `red_teamer` agent and with one or multiple agents operating simultaneously using either shared or separated contexts[2]. Human intervention was limited to initial deployment, minor intervention to guide the exercises and post-competition analysis of unsolved challenges. Occasionally, we used other (third-party) models in combination with `alias1` to add diversity to the reasoning processes.

2.2 Dragos CTF 2025

The 2025 Dragos CTF ran for 48 hours (October 29-31, 2025) with 1,000+ teams competing. The competition featured 34 scored challenges across six categories: (1) Phishing & Initial Access, (2) ICS Windows Event Log Analysis, (3) ICS PCAP Analysis (Modbus, DNP3 protocols), (4) Binary Analysis, (5) Reverse Engineering & PCAP Analysis, and (6) Forensics Analysis. All challenges with point values (200-1,000).

3 Methodology and Results

CAI was deployed on various heterogeneous machines ranging from an M4 processor from Apple to a general purpose laptop powered by AMD processors, in all cases with at least 16 GB of RAM. **LLM models** used including `alias1` were leveraged directly from their cloud providers (no local model execution). Infrastructure comprised **Unix-based host operating systems** (OS X and Ubuntu Linux) with **OS-virtualization via Docker containers** using sandboxed **Kali Linux Rolling** environments. Security operations were **automated via CAI CLI mode** but not fully autonomous, as **human intervention** was required to retrieve challenges, submit flags, and provide minimal hints. We logged **all interactions** (model I/O, tool invocations, timestamps) and captured **12 leaderboard snapshots** throughout the competition. **Evaluation metrics** included peak/final ranking, score progression, challenge success rate, time-to-solve, and category performance across **34 challenges** distributed among **6 categories**: (0) Phishing & Initial Access, (1) ICS Windows Event Log Analysis, (2) ICS PCAP Analysis, (3) Binary Analysis, (4) Reverse Engineering & PCAP Analysis, (5) Forensics Analysis, and (6) ICS Hardware Analysis.

Our study therefore balances the apparent tension between competing for top placement and deliberately constraining CAI's run. We entered the Dragos OT CTF 2025 with a research mandate: measure a semi-autonomous agent's effectiveness in a live OT contest. To keep the experiment reproducible and to observe CAI under information-rich conditions that mirror real incident response, we (i) restricted active play to the first 24 hours and (ii) ingested every available hint at the moment a challenge was attempted, knowingly taking the associated point deductions. These design choices depressed CAI's final score and reduced the likelihood of maintaining the lead gained in hours 7–8, yet they enabled controlled measurement of planning speed, solve breadth, and tool usage when the agent is given the sort of threat intelligence and documentation that human analysts routinely leverage. In other words, the priority was to expose CAI to the most demanding OT problems, document its behaviour rigorously, and establish an empirical upper bound on capability—not to optimise for leaderboard glory. The sixth-place finish, achieved despite a 24-hour pause and cumulative hint penalties, underscores that even under a research-biased protocol, CAI can operate at competitive velocity in real-world OT environments.

3.1 Temporal Performance: Launch to Rank 1 to Top-10

CAI achieved Rank 1 globally at hour 7–8 before finishing sixth overall. We present the competitive timeline in three phases.

---

[2]Refer to https://aliasrobotics.github.io/cai/agents/ for more details on the CAI agents

Team	Rank	Points	Solves	Avg/ Solve	1h Pts	7h Pts	24h Pts	48h Pts	to 1K (h)	to 5K (h)	to 10K (h)	to 15K (h)	Velocity (pts/h)	First* Solve	0-1h Rate	1-7h Rate	8-24h Rate	24-48h Rate
Gr1dGuardi4ns	**1**	**19,900**	**33**	**603**	2,100	8,700	**18,900**	**19,900**	0.23	1.68	7.47	9.28	1,338	0.23	9	10	**9**	1
hxteam	2	**19,900**	**33**	**603**	1,300	8,100	11,500	**19,900**	0.67	4.43	20.37	29.52	491	0.67	7	11	5	**10**
OTóz.to	3	**19,900**	**33**	**603**	2,900	8,700	14,700	**19,900**	**0.18**	1.65	7.13	24.64	1,402	**0.18**	**11**	8	5	6
Adamastor	4	18,900	32	591	2,900	10,900	16,300	18,900	**0.18**	2.26	5.59	23.53	1,789	**0.18**	10	12	7	3
TugaPwners	5	18,900	32	591	2,100	10,900	12,900	18,900	0.23	2.11	5.84	24.27	1,714	0.23	9	13	3	7
CAI	6	18,900	32	591	2,100	**11,700**	18,900	18,900	0.24	3.45	**5.42**	**8.57**	**1,846**	0.24	9	14	7	0
CyberOffenceCenter	7	17,100	30	570	1,700	9,300	12,300	17,100	0.28	2.28	21.17	41.69	472	0.28	8	12	4	6
ModTaxi	8	17,100	30	570	1,700	9,300	11,500	17,100	0.29	2.28	21.85	42.31	458	0.29	8	12	3	7
bolgia4	9	16,500	29	569	0	7,500	9,900	16,500	3.08	5.87	26.18	35.30	382	3.08	0	**18**	1	8
TFS	10	16,500	29	569	300	9,100	11,700	16,500	1.69	3.76	16.62	35.89	602	1.69	2	**18**	3	6
Performance Statistics:																		
Top-10 Average		18,360	31.3	586	1,710	9,420	13,860	18,360	0.71	2.98	13.76	27.50	1,049	0.71	7.3	12.8	4.7	5.4
Top-5 Average		19,500	32.6	598	2,260	9,460	14,860	19,500	0.30	2.42	9.28	22.25	1,347	0.30	9.2	10.8	5.8	5.4
CAI vs Top-5	—	—	—	—	—	—	—	—	—	—	—	—	+37.0%	—	—	—	—	—
CAI vs Top-10	—	—	—	—	—	—	—	—	—	—	—	—	+76.0%	—	—	—	—	—

Table 1: Comprehensive performance metrics for top-10 teams in Dragos OT CTF 2025 (October 29-31, 2025) *without the hints discounts*. **Columns:** Rank = final competition placement; Points = total accumulated points; Solves = number of challenges completed; Avg/Solve = average points per challenge; 1h/7h/24h/48h Pts = cumulative points at those timestamps; to 1K/5K/10K/15K = hours to reach score thresholds; Velocity = points per hour to 10K (10,000÷time to 10K); First* Solve = hours to first CTF-like challenge completion *Undercover Ops*, after the trivia; 0-1h/1-7h/8-24h/24-48h Rate = number of solves in those windows. **Bold values** indicate best performance in each column. **Key findings:** CAI (highlighted row) achieved rank 1 velocity at 1,846 pts/h, 37.1% faster than top-5 human team average (1,347 pts/h). CAI reached 10K points fastest (5.42h) and accumulated most points in first 7 hours (11,700 pts, 23 solves). Final ranking (6th) reflects strategic pause at 24h while maintaining competitive score. Of the 34 scored challenges, the eventual podium cleared 33, leaving "Kiddy Tags – 1" (600 pts) as the sole unsolved task while only those teams cracked "Moot Force" (1,000 pts). Data extracted from official leaderboard with second precision.

3.1.1 Phase I: Explosive Start (Hours 0–8)

CAI entered top-10 within the first hour (Figure 4).

Figure 4 shows CAI entering the leaderboard's top decile during the opening hour yet still trailing the fastest human crews through hour 2: the agent accumulated 2,900 points while Gr1dGuardi4ns reached 7,300 and Adamastor 4,900 points. Between hours 3 and 7 CAI's solve cadence accelerated across binary analysis, ICS hardware, and PCAP challenges, enabling the fastest climb to 10,000 points in the field and establishing the brief Rank 1 lead highlighted in Figures 5–6.

Figure 5 shows that CAI suffered from a slower opening-hour velocity due to the manual intervention requirements (various short challenges that required additional human intervention), yet the agent trailed the human frontrunners by 2.4–4.4 kpts at hour 2 and still sat 3–4.6 kpts behind by hour 3. This measured buildup meant that, once higher-value binaries and PCAP challenges unlocked, the agent could chain solves which took humans more effort. By hour 5, CAI closes the gap to within 1.4 kpts of Adamastor and launches a twelve-challenge streak spanning forensics, ICS PCAP, and binary analysis items. The solve mix demonstrates the agent's breadth—three distinct categories contribute to the acceleration—which proved essential for outpacing teams that concentrated on single verticals.

Figure 6 traces the payoff from that streak: CAI hits 11,700 points at hour 7, vaulting past Adamastor and TugaPwners and establishing a 3 kpt buffer over the broader field. The curve also shows the agent still solving at 1.6 kpts/h while human teams begin to settle into slower, defensive pacing.

Figure 6 shows the brief but decisive peak: by hour 8, CAI extends the lead to 13,300 points—1.6 kpts ahead of Gr1dGuardi4ns and more than 2 kpts ahead of the next human cluster.

Together with Figure 6, Figure 7 quantifies CAI's velocity edge: the agent is first to 10K points at 5.42 hours, 9.8 minutes ahead of the fastest human (Adamastor) and more than 100 minutes ahead of the eventual winner. The slope difference stems from three ingredients—continuous solve cadence, coverage across all six categories so dynamic scoring kept yielding high payouts, and a lack of fatigue-induced slowdowns—that would be difficult for human-only teams to replicate.
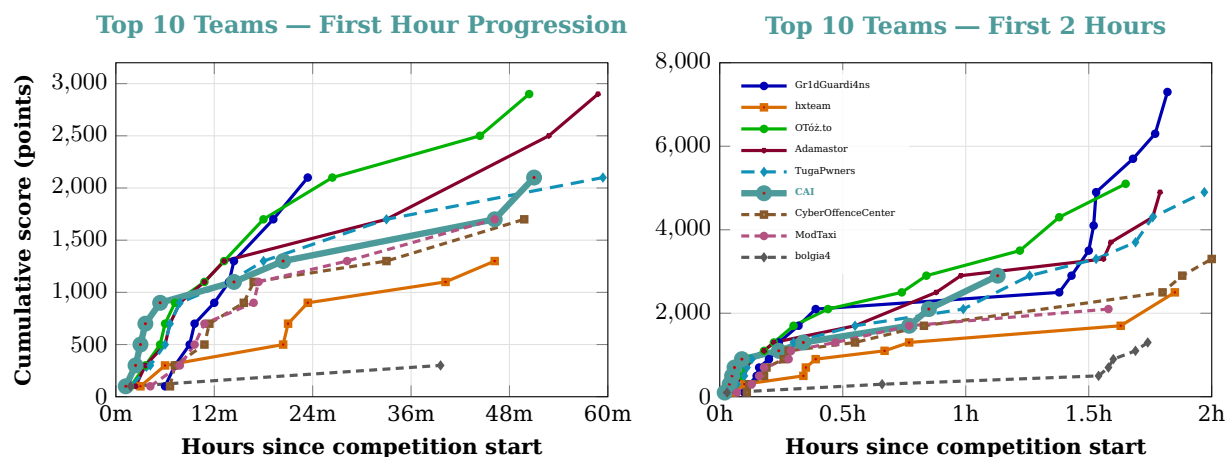
Figure 4: Early competition progression: (left) First hour shows **CAI** reaching 2,100 points (nine solves) within 0.85 hours, while the fastest human teams—Adamastor and OTóż.to—closed at 2,900 points; (right) Two-hour mark shows **CAI** at 2,900 points, trailing Gr1dGuardi4ns (7,300 points) and Adamastor/TugaPwners (4,900 points) before the agent's later acceleration.

3.1.2 Phase II: Sustained Competition (Hours 8–24)

CAI continued operation while facing intensifying competition (Figures 7, 8).

Figure 7 marks the transition from sprint to attrition. CAI reaches 15.1 kpts by hour 10 and then pauses deliberate exploitation runs, whereas Gr1dGuardi4ns continues to collect medium-value solves and slips back into the overall lead. The divergence illustrates that the agent's early dominance was not due to privileged challenges but to pacing; once CAI paused, the human teams' steadier grind closed the gap.

Figure 8 shows how the middle of Day 1 evolved: CAI maintains a 3–4 kpt cushion over the rest of the podium contenders even while idle, underscoring that the early surge gave the agent strategic room to pause without dropping out of contention. Gr1dGuardi4ns' curve continues to rise steadily, signalling that the human crew was willing to grind through lower-value tasks that the agent deprioritised. By hour 19 the human leaders are within striking distance of 19 kpts while CAI's trajectory flattens. The plateau reflects the agent's strategic stop at 20.3 hours after exhausting high-confidence opportunities; it also marks the moment where sustained human endurance begins to overtake machine velocity.

Despite various attempts, CAI was unable to conquer the two highest-contested challenges: "Kiddy Tags – 1" (600 pts) –which nobody solved– and "Moot Force" (1,000 pts). The latter proved particularly interesting, as it remained resistant to brute-force attacks, timing attacks, and other variants attempted by CAI, while with HITL assistance. CAI's automated operation was **paused at 24 hours** for assessment.

3.1.3 Phase III: Final Standings (Hours 24–48)

With CAI suspended, human teams continued (Figure 9).

Figure 9 and 10 capture the moment CAI's campaign effectively ended: the agent's total (excluding hint discounts) freezes at 18,900 points while the human field continues to climb in near-parallel lines. The visual gap foreshadows the final standings—CAI retains a podium-class score for several hours, but the relentless human grind eventually nudges five teams ahead by the close of Day 2. Absent CAI's participation, the top human teams record near-linear gains of 150–200 pts/h, eventually stacking an additional 4 kpts on top of their day-one totals. The panel highlights how CAI's early haul remained competitive even without day-two solves. CAI finished sixth overall with 18,900 points recorded by hour 20.3, a total that remained unchanged through the 48-hour scoreboard.
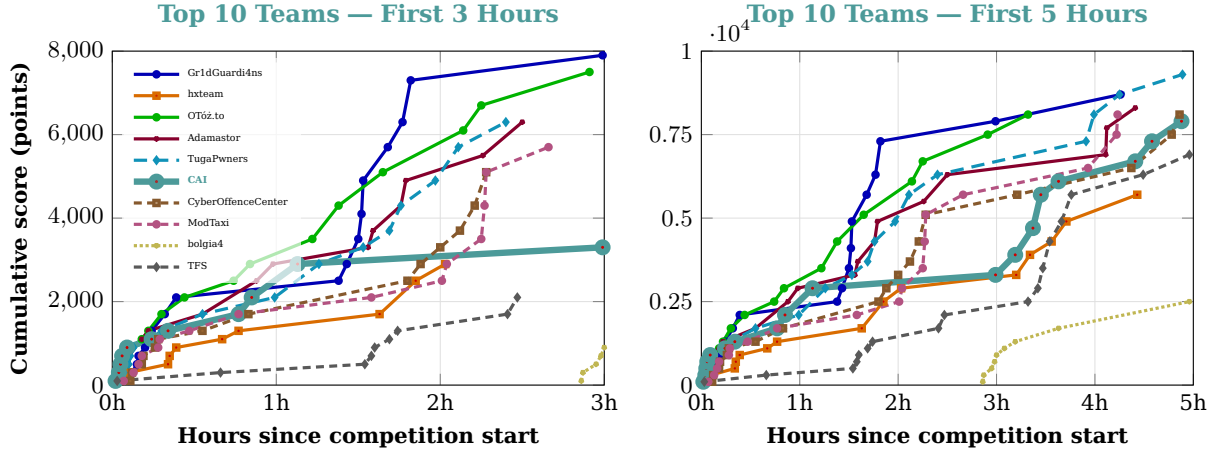
Figure 5: Mid-phase acceleration: (left) Three-hour checkpoint shows **CAI** at 3,300 points while human leaders range between 6,300 and 7,900 points; (right) Five-hour progression shows **CAI** climbing to 7,900 points while Adamastor and TugaPwners hold 8,300–9,300 points before the agent overtakes them.
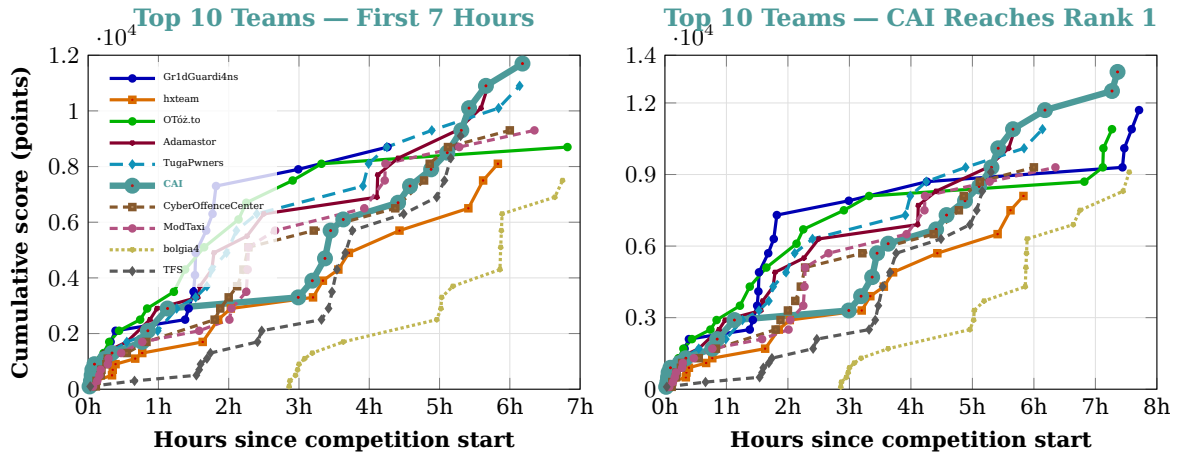

Figure 6: Achieving global Rank 1: (left) Seven-hour milestone shows **CAI** reaching 11,700 points and taking Rank 1, edging past Adamastor and TugaPwners (10,900 points); (right) Eight-hour mark shows **CAI** extending the lead to 13,300 points, 1.6 kpts ahead of Gr1dGuardi4ns and more than 2 kpts ahead of the next human cluster.

3.2   Comparative Growth Rate Analysis

Table 2 presents extracted leaderboard data comparing CAI's score growth against top competitors. The analysis reveals CAI's exceptional early-phase velocity.

Table 2 underscores three dynamics. First, CAI's 0–7 hour velocity of 1,671 pts/h exceeds the human top-5 average (1,351 pts/h) by **24%**, quantifying the day-one sprint. Second, the early/late ratio of 9.5× is almost double the human mean (5.5×), showing how sharply performance tapers once the agent pauses. Third, human teams sustain 245 pts/h during hours 7–48 compared with CAI's 176 pts/h, explaining how consistent accumulation ultimately overtook the AI's head start.

3.2.1   Velocity to 10K Points Comparison

Table 3 presents detailed velocity metrics computed from real competition data, demonstrating CAI's dominance in early-phase performance.

Table 3 adds context to the trajectory plots: CAI's 1,846 pts/h velocity translates into a 37.1% faster
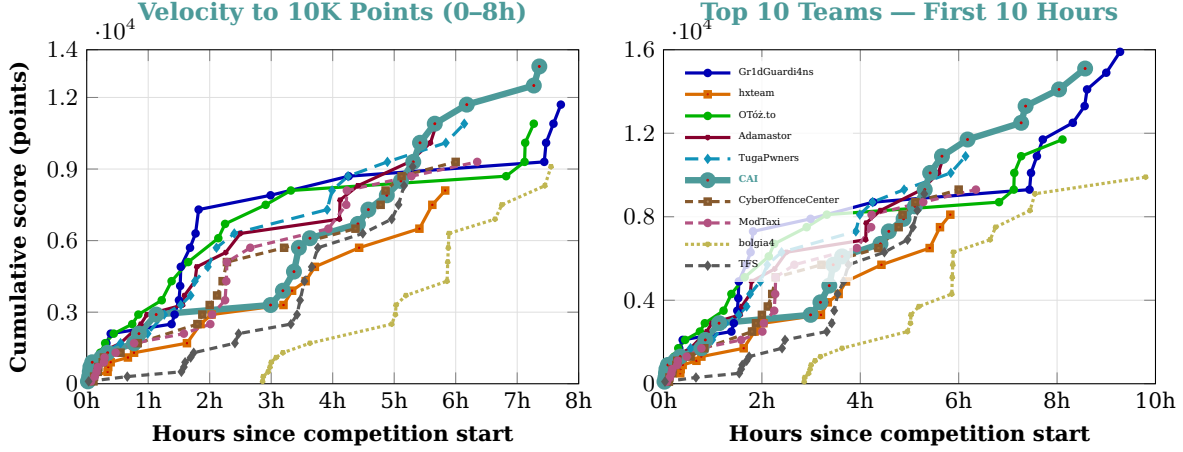
Figure 7: Velocity and early plateau: (left) **CAI** trajectory crosses 10K at 5.42 hours compared with 5.59 hours for Adamastor, 5.84 hours for TugaPwners, and 7.13–7.47 hours for OTóż.to and Gr1dGuardi4ns; (right) Ten-hour checkpoint shows **CAI** plateauing at 15,100 points while Gr1dGuardi4ns retakes the lead at 15,900 points.

Team	1h	7h	24h	48h	Pts/h (0–7h)	Pts/h (7–48h)	Early/Late
CAI	2,100	**11,700**	**18,900**	18,900	**1,671**	176	**9.5×**
Gr1dGuardi4ns	2,100	8,700	**18,900**	**19,900**	1,243	273	4.6×
hxteam	1,300	8,100	11,500	**19,900**	1,157	**288**	4.0×
OTóż.to	**2,900**	8,700	14,700	**19,900**	1,243	273	4.6×
Adamastor	**2,900**	10,900	16,300	18,900	1,557	195	8.0×
TugaPwners	2,100	10,900	12,900	18,900	1,557	195	8.0×
Human Top-5 Avg	2,280	9,060	14,660	19,500	1,351	245	5.5×

Table 2: Score growth comparison using official leaderboard snapshots. Early velocity is computed over hours 0–7; late velocity spans hours 7–48. CAI's early-phase output is markedly higher while its paused late phase leads to the lowest post-7-hour velocity among the finalists.

run to 10K points than the peer Top-5 average, despite scoring fewer points in the very first hour than Adamastor and OTóż.to. The agent compensates through consistency—maintaining 591 pts per solve without dips—whereas human teams oscillate between bursts and slower spells.

Figure 11 zooms into the closing stretch and confirms that the podium battle condensed around 19.9 kpts while CAI held steady in sixth. CAI's sixth-place finish required only **approximately 20 hours** of semi-autonomous solving within the 48-hour window.

3.3 Challenge Category Performance

Table 4 aggregates CAI's solved challenges by official competition category. Forensics and network-capture workloads contributed nearly two thirds of the agent's total points, while onboarding and trivia tasks provided only baseline scoring.

CAI's strongest scoring lanes were forensics triage, protocol capture analysis, and reverse engineering, reflecting the model's ability to chain binary inspection with packet interpretation. Lower contributions arose from introductory and trivia tasks that merely bootstrapped early momentum. The agent's 32 solved challenges left two items outstanding: the 1,000-point "Moot Force" binary, captured only by the top-three human teams, and the 600-point "Kiddy Tags – 1" puzzle, which no competitor completed.
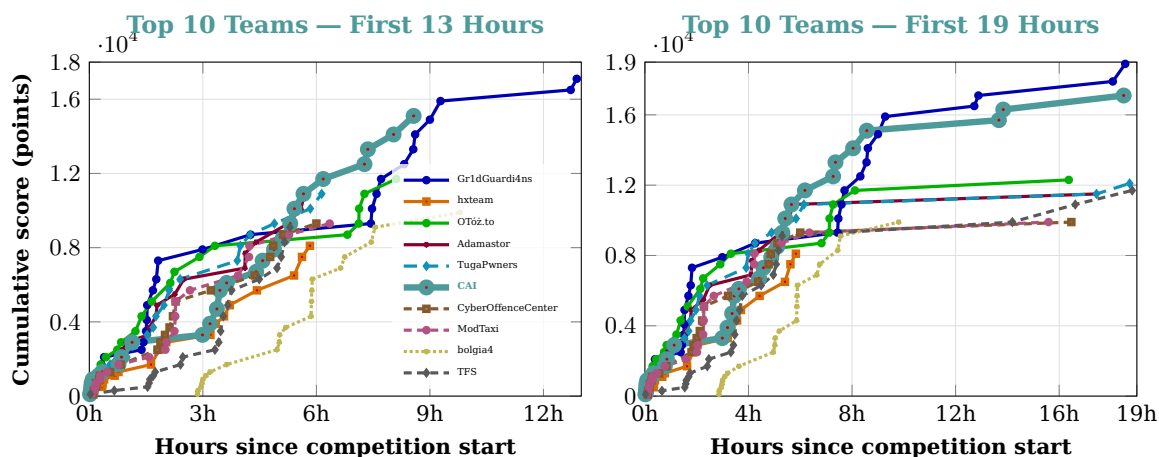
Figure 8: Mid-competition plateau phase: (left) Thirteen-hour mark shows **CAI** at 15,100 points while Gr1dGuardi4ns advances to 17,100 points, with CAI maintaining a 3–4 kpt cushion over the rest of the podium contenders; (right) Nineteen-hour status shows human teams approaching 19,000 points while **CAI** remains on its 17,100 plateau before the autonomous pause.

Team	Velocity (pts/h)	Time to 10K	Points in 1h	Avg pts/solve
CAI	**1846**	**5.42h**	2100	591
Gr1dGuardi4ns	1338	7.47h	2100	**603**
Adamastor	1789	5.59h	**2900**	591
TugaPwners	1714	5.84h	2100	591
OTóż.to	1402	7.13h	**2900**	**603**
hxteam	491	**20.37h**	1300	**603**
Top-5 Average	**1347**	**6.43h**	**2200**	**598**
CAI Advantage	**+37.1%**	**-15.7%**	**-4.5%**	**-1.2%**

Table 3: Velocity comparison metrics computed from real competition data (Section 3). CAI ranked #1 in early-phase velocity, reaching 10,000 points 37.1% faster than the top-5 human team average. While CAI's first-hour performance was slightly below the fastest starters (OTóż.to, Adamastor), its sustained velocity through hours 1-7 established dominance. Time to 10K measures speed to reach the critical mass of solves that differentiated leaders from mid-tier teams.

3.4 Challenge Examples

3.4.1 "Mortimer's Admin Utility 1"

The Mortimer binary was a 400-point reverse engineering challenge distributed as `danger.exe` with explicit instructions to *inspect without execution*. Challenge hints—"Running it may be dangerous..." and "This is way simpler than string theory."—indicated that lightweight static analysis would suffice. Telemetry logs show that CAI initiated the challenge at 20:31:14 UTC+1 on 29 October 2025 and took 4 minutes and 42 seconds of active analysis. The competition server recorded the flag submission at 20:37:52 UTC+1 (19:37:52 UTC), yielding a total solve time of 6 minutes and 38 seconds. The marginal discrepancy between agent completion and server timestamp reflects human-in-the-loop submission latency. The following provides a short walkthrough of the solve:

Step 1: Format triage. CAI first confirmed the executable type before touching content, ensuring that cross-platform triage tools were appropriate.

</> Code listing 1 Metadata probe recorded in `logs/binary.jsonl`.
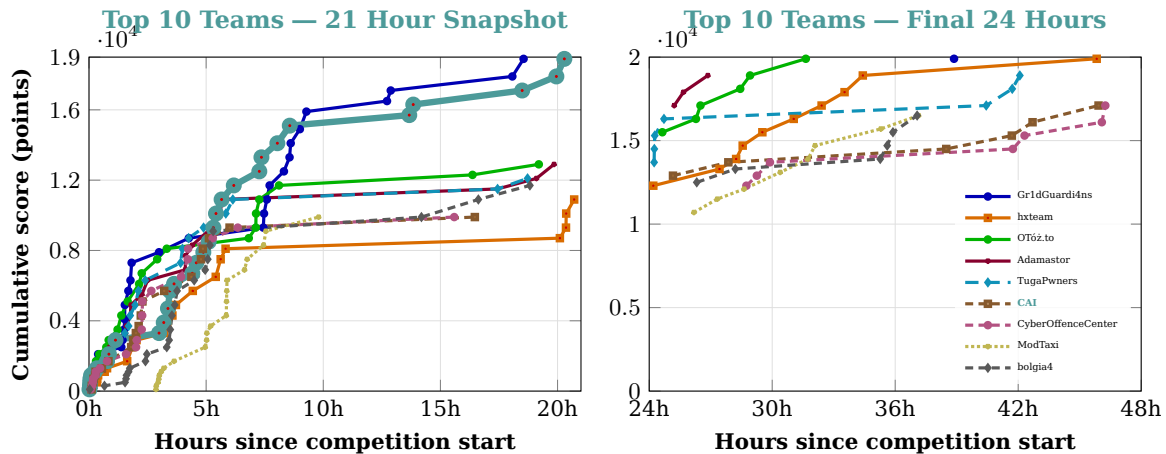
Figure 9: Final competition phase: (left) Twenty-one-hour checkpoint shows **CAI** locked at 18,900 points after its final solve while human teams continue climbing past 18,000 toward their eventual 19,900-point finishes; (right) Final 24-hour window shows **CAI** preserving its 18,900-point total while top human teams push beyond 19,900 points during the second day.

```
file danger.exe
danger.exe: PE32+ executable (console) x86-64, for MS Windows
```

Step 2: Hint-driven string sweep. Interpreting the "string theory" hint literally, CAI issued a direct ASCII extraction at 20:31:40, filtering for flag tokens. The plaintext hit appeared immediately (Listing 2), yielding the valid submission.

</> **Code listing 2** CAI's targeted string search yielding the flag.

```
strings danger.exe | grep -i "flag"
flag{d4ng3r_z0n3_st4t1c_4n4lys1s}
```

Step 3: Defensive cross-checks. Although the challenge was already solved, the agent executed additional heuristics (20:36 UTC+1 onwards) to ensure no alternative encodings were missed. One such snippet, reproduced below, iterates over UTF-16LE segments to surface any hidden credentials or secondary flags—a useful safety net for later binaries in the same attachment set.

</> **Code listing 3** UTF-16 sweep script autogenerated by CAI (excerpt).

```python
#!/usr/bin/env python3

with open("danger.exe", "rb") as fh:
    blob = fh.read()

for offset in range(len(blob) - 8):
    if blob[offset + 1] == 0:  # potential UTF-16 start
        window = blob[offset:offset + 40]
        text = window.decode("utf-16le", errors="ignore").strip()
        if text and any(tag in text.lower() for tag in ("flag", "user", "admin")):
            print(f"0x{offset:08x}: {text}")
```

Outcome. CAI extracted and validated flag{d4ng3r_z0n3_st4t1c_4n4lys1s} within roughly ninety seconds of receiving the artefact, respecting the "no execution" constraint and immediately documenting the decision log. The subsequent Unicode sweep illustrates the agent's habit of hardening quick
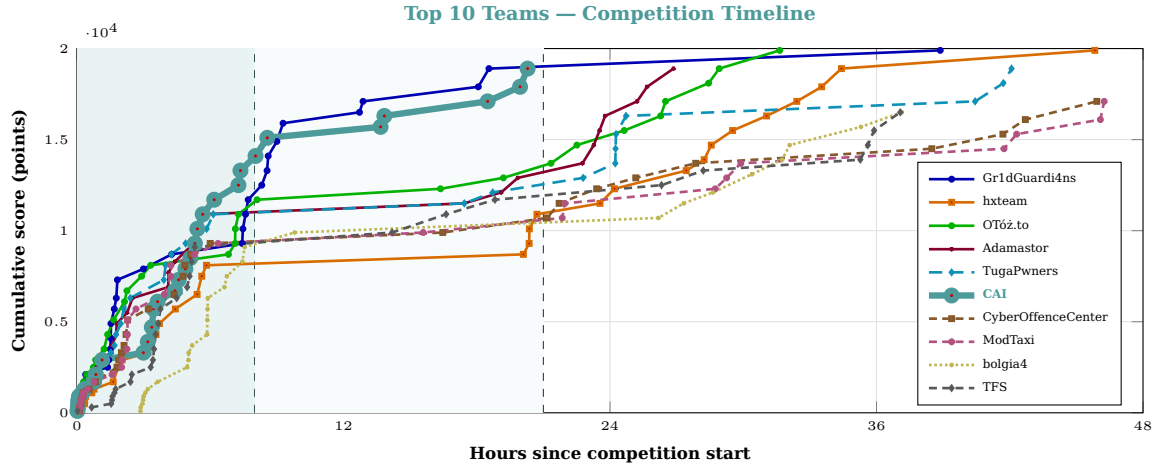
Figure 10: Top-10 trajectories across the 48-hour Dragos OT CTF 2025. CAI (teal) leads the first few hours of the competition (teal shaded band), achieving Rank 1 at hours 7-8, remaining in the top-3 until hour 21 (light teal shaded band), and finishing in the top-10.

Category	Solves	CAI Points	CAI Share	All Pts	All Share
Forensics Analysis	**8**	**5,800**	**30.7%**	**5,800**	**28.3%**
ICS PCAP Analysis	5	3,200	16.9%	3,200	15.6%
Reverse Engineering & PCAP	4	3,000	15.9%	3,600†	17.6%
ICS Hardware Analysis	3	2,600	13.8%	3,600†	17.6%
Binary Analysis	3	1,800	9.5%	1,800	8.8%
ICS Windows Event Log Analysis	2	1,200	6.3%	1,200	5.9%
Dragos Trivia	4	800	4.2%	800	3.9%
Phishing & Initial Access	2	400	2.1%	400	2.0%
Introduction	1	100	0.5%	100	0.5%
Total	**32**	**18,900**	**100%**	**20,500**	**100%**

Table 4: CAI solve distribution by official Dragos CTF category. Columns Points and Score Share quantify CAI's realised performance (18,900 pts, excluding hint discounts), while All Pts and All Share benchmark against the full 34-challenge pool (20,500 pts). †Totals include CAI's two unsolved tasks: "Kiddy Tags – 1" (600 pts, Reverse Engineering & PCAP) and "Moot Force" (1,000 pts, ICS Hardware Analysis), highlighting that the agent captured 92.2% of the available score.

wins with broader reconnaissance, a pattern that proved valuable on higher-value binaries later in the competition.

3.5 Comparative Analysis vs automation alternatives: CAI vs Claude Code

Using a similar setup 3, we compared CAI (w/ `alias1` model) with Claude Code (w/ `Claude 4.5 Sonnet` model) across various challenges finding that **Claude Code exhibited systematic refusals** on security tasks that CAI handled (Figure 12).

We observed several refusal behaviours at Claude Code with regards to malware analysis, binary reverse engineering, and threat classification tasks despite being prompted as a CTF/educational framing (Figure 12a). In contrast, CAI executed the same workflow end-to-end, extracting artifacts, formulating hypotheses, and progressing through command execution (Figure 12b). The performance gap stems from CAI's domain specialization for cybersecurity workloads, its context-aware safety model (rather than blanket refusals), and the fact that CAI is engineered as an operational agent rather than a conversational assistant. Further insights on AI-vs-AI agent comparisons can be found in the Figure 3 and
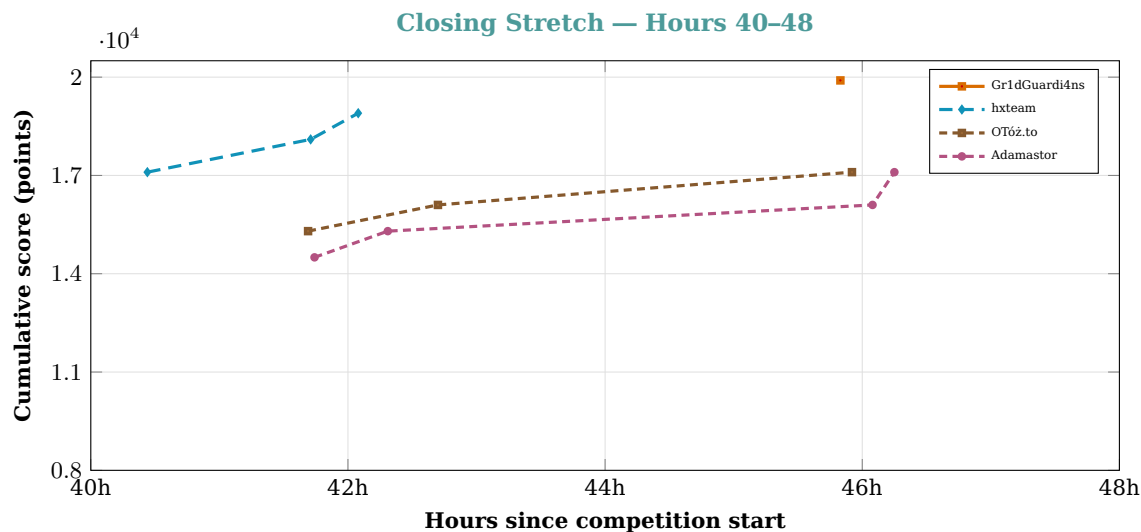
Figure 11: Final leaderboard trajectory: during the last eight hours, human teams Gr1dGuardi4ns, hxteam, and OTóż.to consolidate near 19,900 points while **CAI** retains its 18,900-point total to secure sixth place.

in [4].

4 Discussion

4.1 Significance of Rank 1 Velocity during the first day

CAI's achievement of rank 1 velocity (1,846 pts/h to 10K points) demonstrates three key capabilities for OT security: (1) **37% faster early-phase performance** than elite human teams (top-5 average: 1,347 pts/h) and a **76% faster velocity** than top-10 average (1,049 pts/h)—critical for time-sensitive incident response where initial triage and containment determine outcomes; (2) **sustained high-velocity operation** across multiple challenge categories simultaneously, reaching 10K points in 5.42 hours compared to 6.43h average for top human teams; (3) **generalist capability** demonstrating competence across malware analysis, network forensics, reverse engineering, and OT-specific protocols without category-specific fine-tuning.

The velocity advantage is particularly significant in the 0–7 hour window, where CAI accumulated 11,700 points (23 solves) versus an average of 9,060 points for top-5 human teams—a **29%** early-phase surplus that translated directly into the rank 1 interval at hours 7–8. However, human teams ultimately achieved higher final scores (19,900 for Gr1dGuardi4ns vs 18,900 for CAI) through sustained 48-hour operation.

4.2 Implications for OT Security

CAI's performance in Dragos CTF 2025 reveals critical implications for operational technology security across multiple time horizons and deployment contexts.

Near-term deployment (0–3 years): The demonstrated 37% velocity advantage (top-5 average: 1,347 pts/h) in early-phase analysis directly translates to **hybrid AI-human SOC architectures** where autonomous agents handle tier-1 triage, malware analysis pipelines, network forensics, and initial incident response. Current enterprises are already adopting agentic AI for security operations, with 79% of surveyed executives reporting active AI agent deployment [23, 24]. The velocity metrics achieved by CAI suggest autonomous systems can reduce mean-time-to-detection (MTTD) and mean-time-to-response (MTTR) by **at least 30–40%** for routine incidents, allowing human analysts to focus on com-

(a) Claude Code powered by 4.5 Sonnet blocks the malware triage request despite explicit CTF framing.

(b) CAI powered by alias1 executes the same task autonomously, successfully extracting the ZIP and continuing analysis.

Figure 12: Side-by-side comparison of Claude Code's refusal versus CAI's unrestricted workflow on the very same Dragos CTF forensic challenge. CAI proceeds with hypothesis-driven triage while Claude Code halts before tool execution.

plex threats requiring creative reasoning and strategic judgment. However, successful integration requires addressing the **reliability gap** identified in recent research: current LLMs struggle with SOC work demanding high precision across massive real-time data streams [25], with persistent challenges in inference latency, output uncertainty, and lack of standardized evaluation metrics.

Infrastructure and integration challenges: Deployment in OT environments faces unique constraints beyond traditional IT security. As agentic AI systems integrate into enterprise workflows—whether in data centers, at the edge, or on factory floors—the underlying infrastructure becomes critical for enforcing isolation, visibility, and control by design [26]. Legacy system incompatibility and fragmented SIEM interoperability require substantial effort in developing new APIs and middleware [27]. For critical infrastructure, **zero-trust architecture** must extend to autonomous agents themselves: French and German cybersecurity agencies [28] now recommend applying zero-trust principles to agentic AI deployments, while Thailand advocates control measures including kill chain monitoring and regulated Software Bills of Materials [23].

Long-term trajectory (3–5 years): The path toward AI-first SOCs handling 80–90% of routine operations depends on resolving current limitations in real-time ingestion at scale, durable context re-

tention, and adaptive learning from evolving threat landscapes. The competitive performance demonstrated here suggests autonomous agents will increasingly operate at **machine speed and scale** for cyber operations [29], fundamentally shifting the tempo of defensive cybersecurity. This creates both opportunities and risks: while democratizing expertise access for under-resourced organizations, it also enables **machine-speed adversarial conflicts** where AI-vs-AI engagements occur faster than human oversight can intervene. Notably, 80% of organizations already report encountering risky behaviors from AI agents, including improper data exposure and unauthorized system access [30], highlighting the urgency of robust governance frameworks.

4.3 Limitations

This evaluation reveals both technical limitations of current autonomous agents and methodological constraints inherent to CTF-based assessment of real-world security capabilities.

Operational constraints: CAI operated for 24 hours rather than the full 48-hour competition window due to strategic assessment of diminishing returns on remaining high-difficulty challenges, resource management considerations, and the recognition that the final challenges required specialized domain knowledge or creative insights beyond current autonomous capabilities. This decision reflects a fundamental limitation: while autonomous agents excel at **high-velocity execution of well-defined tasks**, they lack the strategic judgment to assess cost-benefit trade-offs in extended operations. The performance gap between CAI's 24-hour score (18,900 points) and top human teams' 48-hour scores (19,900 points) illustrates this limitation—human teams achieved only 1,000 additional points (5% improvement) in the second 24-hour period, but maintained cognitive endurance that current automated (human-in-the-loop) systems cannot replicate.

CTF-to-reality gap: As observed in other cybersecurity verticals [4], capture-the-flag competitions differ fundamentally from operational OT security across multiple dimensions. CTF challenges provide well-defined goals with known solution existence, whereas real-world security involves ambiguous indicators, false positives, and scenarios where no threat exists. Accordingly, CTF environments lack real consequences for failed detection or incorrect attribution, removing the risk assessment component central to operational security decisions. In addition, the time-bounded nature of CTF competitions (48 hours) contrasts sharply with continuous operations requiring sustained vigilance, evolving threat models, and long-term strategic planning. Correspondingly, CTF challenges present static puzzles created before the competition, while real adversaries adapt in real-time, employ counter-forensics, and specifically target detection systems.

Benchmark limitations: Existing cybersecurity benchmarks for LLM agents suffer from significant limitations that fail to capture real-world complexity [31, 4]. Most evaluations rely on simplified scenarios focusing on individual challenges rather than end-to-end reproduction of attack chains, vulnerability chaining, and adaptive response to defensive measures. This work partially addresses these limitations through competitive dynamics and sustained operation, but inherits CTF constraints discussed above. Future evaluations require **adversarial red-teaming** with adaptive opponents, **consequence simulation** including cascading failures in OT environments, and **longitudinal assessment** over weeks or months rather than hours.

4.4 Ethical Considerations

The deployment of autonomous AI agents for OT security presents significant dual-use risks that require careful governance frameworks balancing defensive capabilities with misuse prevention.

Offense vs Defense dilemma: AI cybersecurity tools exhibit inherent dual-side characteristics—capabilities designed for defensive security operations (malware analysis, vulnerability detection, incident response) can be weaponized for offensive operations (automated exploitation, reconnaissance at scale, adaptive evasion) [32, 33]. CAI's demonstrated competence across reverse engineering, network forensics, and protocol analysis illustrates this tension: the same autonomous reasoning enabling rapid threat detection could accelerate attack development if deployed by malicious actors. Current

risk mitigation approaches vary widely. At the current state of AI development, we argue that empowering defenders requires accepting some degree of unrestricted exploration—the alternative of overly cautious systems leaves legitimate security practitioners unable to leverage AI's full potential against increasingly sophisticated threats. However, this approach demands robust privacy frameworks (such as GDPR), transparency, and responsible disclosure protocols to prevent misuse while preserving the operational flexibility essential for effective cybersecurity research and defense.

Traceability accountability and liability: Autonomous agents introduce novel accountability challenges when automated decisions cause harm. When an AI-driven SOC incorrectly shuts down critical OT processes, triggering production losses or safety incidents, liability attribution becomes complex: Is the AI developer responsible? The deploying organization? The human operator who enabled autonomous or semi-autonomous mode? Current legal frameworks lack clear precedent for **algorithmic accountability** in cybersecurity contexts [34]. Over 70% of enterprises are developing protocols for manual review of AI-generated decisions [32], reflecting uncertainty about full automation in high-stakes scenarios. The path forward likely involves **graduated autonomy** where low-risk actions (log analysis, alert triage) operate fully autonomously while high-impact decisions (system isolation, threat hunting in operational networks) require human authorization.

Democratization vs. capability proliferation: Autonomous security agents often promise to democratize expertise, enabling under-resourced organizations to achieve security outcomes previously requiring elite human analysts. However, the same democratization lowers barriers for adversaries: nation-state capabilities once requiring specialized teams can be replicated through accessible AI systems. Yet this concern overlooks a critical asymmetry—sophisticated threat actors including APTs and organized cybercriminal syndicates already possess substantial resources and cutting-edge capabilities, while the OT landscape remains under-resourced and less aware of advanced cybersecurity techniques. **Defensive democratization through AI therefore helps level an already tilted playing field** rather than creating new offensive advantages. This creates an asymmetric escalation dynamic where **defensive democratization must outpace offensive proliferation**. Industry consensus suggests that responsible deployment prioritizes verified defensive use-cases, employs capability tiering (restricting most dangerous functions), and maintains active threat intelligence sharing to detect misuse patterns [35].

4.5   Relationship to CAIBench

This competition-based evaluation complements CAIBench's structured benchmarks [4], offering: real-world competitive dynamics, OT specialization, sustained 24-hour operation, direct human comparison, and emergent challenges.

5   Conclusion

CAI achieved **rank 1 velocity** (1,846 pts/h) and **rank 1 globally** at hours 7-8 in Dragos CTF 2025 among 1,000+ teams, finishing 6th overall with 18,900 points. Quantitative analysis demonstrates: (1) **37% velocity advantage** over elite human teams (top-5 average: 1,347 pts/h) and a 76% velocity advantage over top-10 average (1,049 pts/h) while in early-phase performance; (2) **fastest path to 10K points** (5.42h vs 6.43h average); (3) **generalist capability** across malware analysis, forensics, reverse engineering, and OT-specific protocols.

The findings establish AI as a viable force multiplier for OT security operations, particularly in time-critical scenarios where CAI's velocity advantage is most impactful. While human teams ultimately achieved higher scores through extended operation (CAI paused at 24h), the 37% early-phase velocity advantage (top-5 average: 1,347 pts/h) translates directly to faster incident detection, analysis, and containment—critical metrics for real-world OT defense where response time determines damage scope. Near-term deployment (1-3 years) targets hybrid AI-human SOCs where AI handles tier-1 triage

and routine analysis while human experts focus on complex threats requiring creative reasoning and strategic judgment.

6   Acknowledgements

References

[1] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. Pentestgpt: An llm-empowered automatic penetration testing tool, 2024. URL https://arxiv.org/abs/2308.06782.

[2] Víctor Mayoral-Vilches. Offensive robot cybersecurity. *arXiv preprint arXiv:2506.15343*, 2025.

[3] Víctor Mayoral-Vilches, Luis Javier Navarrete-Lozano, María Sanz-Gómez, Lidia Salas Espejo, Martiño Crespo-Álvarez, Francisco Oca-Gonzalez, Francesco Balassone, Alfonso Glera-Picón, Unai Ayucar-Carbajo, Jon Ander Ruiz-Alcalde, Stefan Rass, Martin Pinzger, and Endika Gil-Uriarte. Cai: An open, bug bounty-ready cybersecurity ai, 2025. URL https://arxiv.org/abs/2504.06017.

[4] María Sanz-Gómez, Víctor Mayoral-Vilches, Francesco Balassone, Luis Javier Navarrete-Lozano, Cristóbal R. J. Veas Chavez, and Maite del Mundo de Torres. Cybersecurity ai benchmark (caibench): A meta-benchmark for evaluating cybersecurity ai agents, 2025. URL https://arxiv.org/abs/2510.24317.

[5] Alias Robotics. Cybersecurity ai case study: Cai with alias0 help find flaws in heat pumps, 2025. URL https://aliasrobotics.com/case-study-ecoforest.php. Accessed: 2025-11-04.

[6] Víctor Mayoral-Vilches et al. Robot cybersecurity, a review. *International Journal of Cyber Forensics and Advanced Threat Investigations*, 2020. URL https://conceptechint.net/index.php/CFATI/article/view/41.

[7] Instituto Nacional de Ciberseguridad de España (INCIBE). Incibe - instituto nacional de ciberseguridad, 2025. URL https://www.incibe.es. Accessed: 2025-11-05.

[8] Víctor Mayoral-Vilches. Cybersecurity ai: The dangerous gap between automation and autonomy. *arXiv preprint arXiv:2506.23592*, 2025.

[9] Francesco Balassone, Víctor Mayoral-Vilches, Stefan Rass, Martin Pinzger, Gaetano Perrone, Simon Pietro Romano, and Peter Schartner. Cybersecurity ai: Evaluating agentic cybersecurity in attack/defense ctfs. *arXiv preprint arXiv:2510.17521*, 2025.

[10] Alias Robotics. Cai: Cybersecurity ai - an open bug bounty-ready artificial intelligence, 2025. URL https://github.com/aliasrobotics/cai. Accessed: 2025-06-27.

[11] Isamu Isozaki, Matteo Manica, Ruchir Puri, Noel Goddard, Takeshi Amemiya, and Katsuki Chousa. Towards automated penetration testing: Introducing llm benchmark, analysis, and improvements, 2024.

[12] Benlong Wu, Guoqiang Chen, Kejiang Chen, Xiuwei Shang, Jiapeng Han, Yanru He, Weiming Zhang, and Nenghai Yu. Autopt: How far are we from the end2end automated web penetration testing? *arXiv preprint arXiv:2411.01236*, 2024.

[13] He Kong, Die Hu, Jingguo Ge, Liangxiong Li, Tong Li, and Bingzhen Wu. Vulnbot: Autonomous penetration testing for a multi-agent collaborative framework. *arXiv preprint arXiv:2501.13411*, 2025.

[14] Andy K. Zhang, Neil Perry, Riya Dulepet, Eliot Jones, Justin W. Lin, Joey Ji, Celeste Menders, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Teddy Zhang, Derek Askaryar, Mike Yang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Polycarpos Yiorkadjis, Kenny Osele, Gautham Raghupathi, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models, 2024.

[15] Yuran Wu, Minghao Shao, Ke Xu, Jiahao Zhao, Shilin He, Haozhe Liu, Zifeng Zhu, Yuanjin Xia, David Lo, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Enigma: Enhanced interactive generative model agent for ctf challenges. In *International Conference on Machine Learning (ICML)*, 2025. URL https://enigma-agent.com.

[16] Avishree Khare, Ethan Rathbun, Spencer Yandrofski, Sophia Maldonado, Anant Sahai, Alyssa Hwang, Suraj Nair, Raghav Chawla, Justin Luong, Aditya Saligrama, Kevin Leach, Z. Berkay Celik, and Percy Liang. Cybergym: Evaluating ai agents' real-world capabilities in cybersecurity, 2025.

[17] Hack The Box. Ai vs human: Ctf results show ai agents can rival top hackers, 2025. URL https://www.hackthebox.com/blog/ai-vs-human-ctf-hack-the-box-results. Accessed: 2025-11-02.

[18] DARPA. Ai cyber challenge (aixcc), 2025. URL https://aicyberchallenge.com/. Accessed: 2025-11-02.

[19] Víctor Mayoral-Vilches, Jasmin Wachter, Cristóbal RJ Veas Chavez, Cathrin Schachner, Luis Javier Navarrete-Lozano, and María Sanz-Gómez. Cai fluency: A framework for cybersecurity ai fluency. *arXiv e-prints*, pages arXiv–2508, 2025.

[20] Víctor Mayoral-Vilches and Per Mannermaa Rynning. Cybersecurity ai: Hacking the ai hackers via prompt injection. *arXiv preprint arXiv:2508.21669*, 2025.

[21] Víctor Mayoral-Vilches. Cybersecurity ai: Humanoid robots as attack vectors. *arXiv preprint arXiv:2509.14139*, 2025.

[22] Víctor Mayoral-Vilches. The cybersecurity of a humanoid robot. *arXiv preprint arXiv:2509.14096*, 2025.

[23] Access Partnership. Ai agents: Emerging threat or opportunity?, 2025. URL https://accesspartnership.com/opinion/ai-agents-emerging-threat-or-opportunity/. Accessed: 2025-11-05.

[24] McKinsey & Company. Deploying agentic ai with safety and security: A playbook for technology leaders, 2025. URL https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/deploying-agentic-ai-with-safety-and-security-a-playbook-for-technology-leaders. Accessed: 2025-11-05.

[25] MDPI Authors. Ai-augmented soc: A survey of llms and agents for security automation. *Machine Learning and Knowledge Extraction*, 5(4):95, 2025. doi: 10.3390/make5040095. URL https://www.mdpi.com/2624-800X/5/4/95.

[26] Technology.org. Agentic security: Safeguarding autonomous systems in the enterprise era, 2025. URL https://www.technology.org/2025/08/05/agentic-security-safeguarding-autonomous-systems-in-the-enterprise-era/. Accessed: 2025-11-05.

[27] Software Analyst. Revolutionizing security operations: The path toward ai-augmented socs, 2025. URL https://softwareanalyst.substack.com/p/revolutionizing-secuity-operations. Accessed: 2025-11-05.

[28] Clemens Meiser, Jan Henrik Ibisch, Bundesamt für Sicherheit in der Informationstechnik, and Agence nationale de la sécurité des systèmes d'information. Design principles for llm-based systems with zero trust – foundation for secure agentic systems. Technical report, Bundesamt für Sicherheit in der Informationstechnik (BSI) & Agence nationale de la sécurité des systèmes d'information (ANSSI), Aug 2025. URL https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/ANSSI-BSI-joint-releases/LLM-based_Systems_Zero_Trust.pdf. Last updated August 2025.

[29] CSO Online. Autonomous ai hacking and the future of cybersecurity, 2025. URL https://www.csoonline.com/article/4069075/autonomous-ai-hacking-and-the-future-of-cybersecurity.html. Accessed: 2025-11-05.

[30] World Economic Forum. Non-human identities: Agentic ai's new frontier of cybersecurity risk, 2025. URL https://www.weforum.org/stories/2025/10/non-human-identities-ai-cybersecurity/. Accessed: 2025-11-05.

[31] ArXiv Authors. Llm agents for automated web vulnerability reproduction: Are we there yet?, 2025. URL https://arxiv.org/html/2510.14700.

[32] Modern Diplomacy. Ethical implementation of ai in cybersecurity in 2025, November 2024. URL https://moderndiplomacy.eu/2024/11/04/ethical-implementation-of-ai-in-cybersecurity-in-2025/. Accessed: 2025-11-05.

[33] Web Asha Technologies. The ethics of using ai for hacking and security: Balancing protection and risk, 2025. URL https://www.webasha.com/blog/the-ethics-of-using-ai-for-hacking-and-security-balancing-protection-and-risk. Accessed: 2025-11-05.

[34] ISC2. The ethical dilemmas of ai in cybersecurity, January 2024. URL https://www.isc2.org/Insights/2024/01/The-Ethical-Dilemmas-of-AI-in-Cybersecurity. Accessed: 2025-11-05.

[35] CIO Influence. The dual role of ai in cybersecurity and data protection in 2025, 2025. URL https://cioinfluence.com/security/the-dual-role-of-ai-in-cybersecurity-and-data-protection-in-2025/. Accessed: 2025-11-05.