WHITE PAPER

Carnegie Mellon University
CyLab Security and Privacy Institute

Skill or Shortcut?

AI, Competitive Cybersecurity Learning, and the Growing Gap Between Performance and Expertise

Megan Kearns, Luke T. Jones, Ivan Liang, Max Yin

April 2026

About This Paper

Six years of platform data from CyLab's competitive cybersecurity education environment reveals a structural shift in how participants engage with challenges. Beginning in 2023 and accelerating sharply through 2026, performance patterns are no longer consistent with historical variation. This paper documents three findings drawn from population-scale performance data: unprecedented score compression, accelerated solve velocities, and a divergence between entering and returning cohorts. It also reports a fourth, preliminary observation: an emerging behavioral shift toward unsupervised AI agent deployment, characterized here from early and largely qualitative evidence and still under analysis. After isolating these quantitative findings from structural or demographic anomalies, the paper examines what competitive performance actually measures in the presence of capable AI. Finally, it describes how CyLab is evolving its platform to address the growing gap between a learner's ability to obtain an automated solution and their capacity to independently understand, evaluate, and constrain it.

Section 1: The Problem Worth Solving

For the past six years, Carnegie Mellon University's CyLab Security and Privacy Institute has operated one of the most widely used cybersecurity education competitions in the world. picoCTF, now transitioning to CyLab Security Academy, has evolved from a regional student competition into a global learning platform. Internal platform records indicate more than one million registered users, 14K+ classrooms, and over 10.7 million challenge solves across 561 problems spanning major domains of offensive and defensive security. Each year, tens of thousands of students, teachers, and independent learners use the platform to develop practical skills under competitive conditions. In 2026 alone, the competition drew 14,286 participants across 2,978 teams, generating 116,127 challenge solves over a ten-day period. Of those participants, 4,338 were middle or high school students competing alongside university students and independent learners across the open bracket.

That scale is central to this paper because it enables observation of learning behavior over time and across a large population. picoCTF has operated as a competition since 2013, but for most of that history the platform closed between annual events. In late 2020, CyLab transitioned picoCTF to a continuously available platform, retaining the annual competition as an anchor event while making the full challenge library accessible year-round. It is this continuously operating platform, now running for six years, that produces the longitudinal behavioral record this paper draws on: a sustained record of how participants engage with and develop cybersecurity skills in an environment that rewards demonstrated competence. When patterns in that record shift, those changes warrant careful analysis. CyLab occupies an unusual position in relation to the findings this paper presents. The platform that generated the longitudinal behavioral record is the same platform positioned to respond to what that record reveals. The data does not sit apart from the infrastructure. It is produced by it, across six years of continuous operation at a scale no controlled study could replicate. This paper draws on both functions. Sections 1 and 2 report what the observatory observed. Section 3 examines what those observations mean for how competitive cybersecurity education is understood. Section 4 describes how CyLab is responding and remains under active development as the findings in this paper are reviewed and extended. The observational and comparative record presented here stands independently of that response work and is released on that basis.

Beginning in 2023, and accelerating through 2025 into 2026, platform logs suggested measurable changes in participant interaction patterns. Challenges that historically required extended time to solve began resolving more quickly. Categories that had previously seen limited novice success showed increased participation. These observations were not uniform year-over-year variation but directional shifts that warranted formal analysis. The magnitude, timing, and structure of these shifts are inconsistent with historical variation alone. AI-assisted problem solving is the most plausible primary driver.

In 2026, a subset of participants extended beyond conventional tool use and deployed AI-enabled agents against the competition environment. Behavioral logs show automated port scanning, scripted interaction attempts, and high-volume incorrect flag submissions at speeds not consistent with manual human interaction. While some of these actions resulted in disqualification under competition rules, they also provide insight into how participants are experimenting with increasingly autonomous

systems. In several cases, the observed behavior is consistent with participants deploying systems whose actions they could not fully supervise or predict.

This paper does not focus on competition enforcement. Although integrity management is an operational concern, the behaviors associated with AI-enabled misuse provide a window into a broader shift: participants are adopting AI systems rapidly, often ahead of developing a clear understanding of how those systems function or how to constrain them effectively. In cybersecurity contexts, where actions can have unintended consequences beyond the immediate environment, that gap carries practical significance.

This shift is not isolated to a single platform, and over 2025 it moved quickly. An early signal came in April 2025, when Hack The Box ran a controlled AI vs Human competition: five of eight AI-agent teams solved 19 of 20 challenges, a 95 percent solve rate, and the top AI team finished just outside the top 20 among 403 human teams. They were competitive with strong human play, but not yet ahead of the best. Over the rest of the year that gap narrowed. Subsequent research has documented AI systems successfully competing in and, in some cases, outperforming human participants in Jeopardy-style CTF environments (Mayoral-Vilches et al., 2025). Platform operators have begun adapting accordingly, introducing mechanisms that allow organizers to explicitly enable or restrict AI usage within competitions. At the same time, practitioners have increasingly questioned whether traditional formats can continue to measure individual skill acquisition in the presence of capable AI assistance.

This paper draws on internal platform data, including longitudinal participation records and challenge interaction logs, alongside a focused 2026 dataset capturing AI-mediated behaviors in a live competitive security environment. The goal is not to characterize these developments as uniformly negative. Some observed patterns suggest increased access, faster onboarding, and broader participation. However, taken together, the findings point toward a widening gap between the ability to use AI systems and the ability to understand and appropriately supervise them.

Competitive cybersecurity education environments provide a rare setting in which that gap becomes observable, measurable, and testable. CyLab has operated such an environment at scale for several years. This paper examines what those observations reveal.

Section 2: What the Data Shows

Finding 1: The Performance Landscape Shifted — But Not Gradually

For the first four years of this study, competitive performance on picoCTF followed a recognizable pattern. The top teams were clearly separated from the competitive middle. Hard challenges filtered out most competitors. Returning users who competed on the same account across multiple years showed modest improvement. The gap between the best teams and the rest of the field was large and stable. The platform was functioning as intended: differentiating levels of problem-solving ability, separating harder problems from easier ones, and producing measurable variation across the competitive field. picoCTF 2026 competition is not simply a stronger field. It is a structurally different one, where solutions are reached faster, more uniformly, and more broadly distributed across participants than any prior year would have predicted.

The initial pattern held through 2024, though 2025 showed early signs of compression. Then 2026 broke it entirely: median time-to-first-solve for hard challenges, which ranged 40 to 340 minutes across the prior five years, plummeted to 5 minutes in 2026, while hard challenge field-wide solve rates roughly tripled from prior years' 5-7% to above 18%.

The clearest way to see what happened is through the competitive tier, the teams ranked six through fifty on the global leaderboard in each competition year. These are not casual participants but highly engaged competitors who solved most of the available challenges and finished near the top of a global field. From 2021 through 2025, this group solved between 71% and 92% of hard challenges in any given year, with meaningful variation between years. In 2026, every team in the competitive tier solved every hard challenge, eliminating variation in a difficulty tier specifically designed to differentiate performance.

The score compression tells the same story. In 2021, the gap between the top team and the fiftieth-ranked team was substantial, reflecting meaningful differentiation in both skill and speed. By 2025 that gap had narrowed considerably. In 2026 it closed to zero: the top fifty teams all completed every challenge and finished with identical scores.

Final rankings were therefore determined by completion time rather than score, with the highest-ranked teams distinguished only by how quickly they achieved full completion. While this preserves a ranking outcome, it represents a shift in what the competition measures. In a format designed to differentiate competitors based on problem-solving performance, score no longer functioned as a discriminating metric at the top of the field.

This represents not only a performance outcome but a measurement change: when all top competitors achieve identical completion, the competition distinguishes speed of execution rather than depth or breadth of problem-solving ability.

Year	Competitive tier (ranks 6-50) hard solve rate	Score gap rank 1 to rank 50	Teams finishing within 24 hours
2021	71%	5,150 pts	0%
2022	92%	1,000 pts	0%
2023	79%	1,800 pts	4%
2024	88%	1,500 pts	6%
2025	81%	1,500 pts	0%
2026	100%	0 pts	18%

Fig. 1 — Note: 2025 and 2026 competitions ran 10 days. All prior years ran 14 days. The 2022 score gap reflects an expanded challenge set and is not directly comparable to other years on the same scale.

The speed shift reinforces the score data. In 2021, no top-50 team finished the competition within 24 hours. Competitors worked steadily across the full competition window, often returning to challenges over multiple days. In 2026, 18% of the competitive tier completed every challenge in the first day. One team finished in under eight hours. This shift in completion timing indicates that the reduction in score

variation is not only an outcome effect but also a process effect: competitors are reaching complete solutions earlier and more uniformly.

The performance shift is not limited to elite competitors. Among first-time participants, the percentage solving at least one hard challenge more than doubled from 2025 to 2026, reaching a level the platform had not previously seen from participants in their first competition. On average, a first-time participant in 2026 solved four times as many hard challenges as a first-time participant in 2021. Students arriving for the first time are performing at a materially different level than their predecessors on the same set of metrics.

The US middle and high school bracket, which serves as a useful window into what a defined student population is capable of in any given year, showed the same break. Half of all US middle and high school teams solved at least one hard challenge in 2026, up from roughly one in five the year before, representing a substantial shift in the baseline performance of a defined student population. This narrowing of the gap between the US middle and high school bracket and open-bracket performance suggests that access to viable solution pathways has increased faster than the development of underlying expertise. At the top of that bracket, the best teams solved hard challenges at a rate approaching the theoretical maximum. This convergence raises an important measurement question: whether observed performance reflects underlying expertise or increased access to external solution support.

One finding complicates the straightforward interpretation. Among users who returned to compete on the same account in multiple years, those whose second competition fell after 2023 showed significantly lower improvement rates in both score percentile and hard challenge performance than those who returned before AI tools were widely available. This is counterintuitive if AI tools are simply making everyone better. One plausible explanation, developed further in Finding 3, is a ceiling effect: when the field around you improves dramatically, maintaining your relative standing requires more than marginal improvement. Getting better in absolute terms is no longer sufficient to improve your percentile rank. The data suggests a second explanation is also operating, one that Finding 3 examines in detail.

What the data in Finding 1 documents is that a structural shift occurred across multiple independent measures of performance. What it cannot determine on its own is the mechanism by which that shift occurred. The evidence bearing on that question is taken up in Findings 2 and 3 and weighed in Section 3; Finding 4 describes a related behavioral mode rather than the mechanism behind the performance shift itself.

In 2026, the competition did not simply become easier or more competitive. It became less able to distinguish differences in performance at the top of the field.

Finding 2: The Competition Is Being Solved Faster — Much Faster

Performance scores tell you what competitors accomplished. Timestamps capture a different dimension of behavior: how competitors progressed through problems over time. When a competitor solves a hard challenge, when they do it relative to the competition opening, and how that pattern distributes across the field are behavioral signals that scores alone cannot capture. Across six years of data, those signals show a consistent and accelerating shift toward faster, more front-loaded solving behavior that reached a distinct break point in 2026.

In 2021, the mean time of first solves across all hard challenges was 309 minutes after the competition opened. Competitors were working through problems methodically over hours. In 2026 it dropped to 29 minutes. The median time to first solve dropped from around 40 minutes in 2024 and 2025, to just 5 minutes in 2026. The acceleration was not evenly distributed across the hard challenge set; it was concentrated at the front of the window, with most challenges resolving in the opening minutes of the competition.

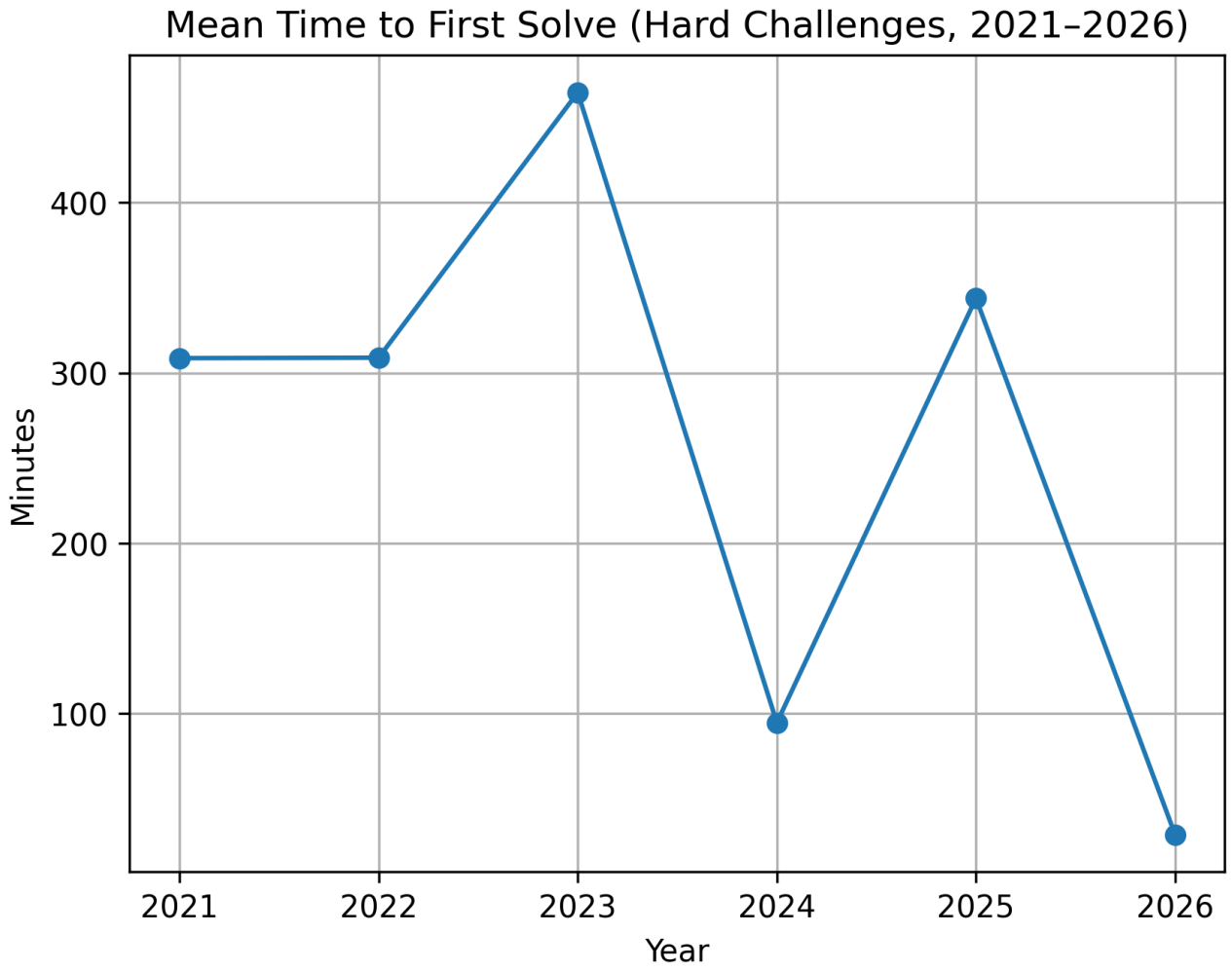

Fig. 2 — Mean time to first hard challenge solve, by year (minutes). As shown, this reduction represents discontinuity relative to prior years rather than a gradual trend.

Hard challenges that would have taken a typical competitor the better part of a day to crack in prior years were being solved within the first half hour of the competition opening. This acceleration is accompanied by increased success rates on the same class of problems (Figure 3), indicating that faster solutions are not isolated early solves but are distributed across a larger portion of the field.

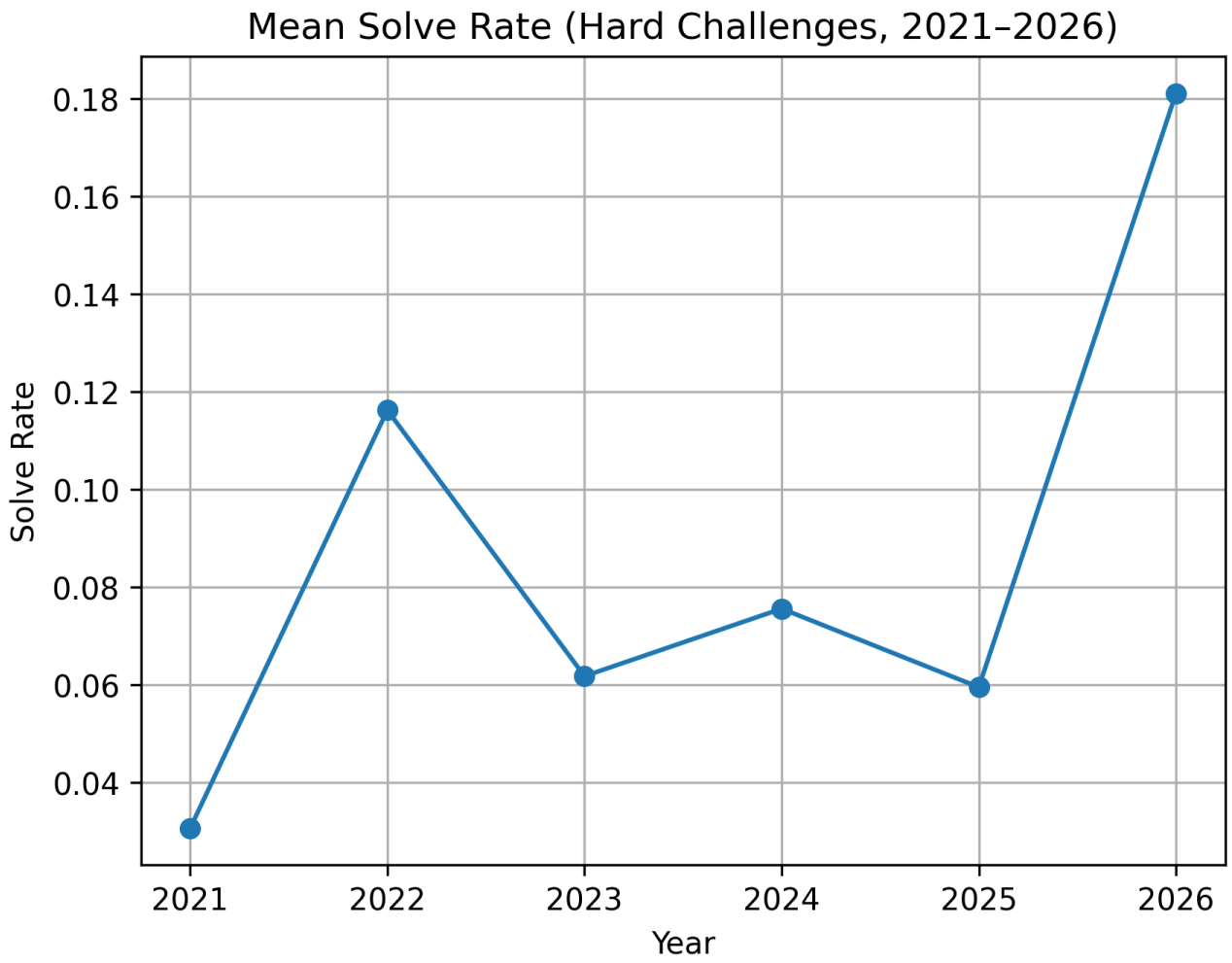

Fig. 3 — Mean hard challenge solve rate, by year. The sharp increase in 2026 accompanies the velocity shift documented in Figure 2.

The same pattern appears in how solves distribute across the competition window. Top-50 teams have always tended to front-load their solving, but the degree intensified sharply. In 2021, 88% of top-50 team solves occurred in the first half of the competition window. By 2026 that figure reached 99%. The back half of the competition, historically where competitors returned to challenges they had left unsolved, became nearly irrelevant for the top tier. They were completing the competition before the back half of the competition window meaningfully contributed to performance differentiation.

Across the broader field, the percentage of all challenge solves occurring within the first 24 hours rose meaningfully from 2023 onward (Figure 4). Figure 4 shows a marked increase in early solve concentration, particularly for hard challenges, indicating a shift from extended, iterative problem-solving to rapid early convergence.

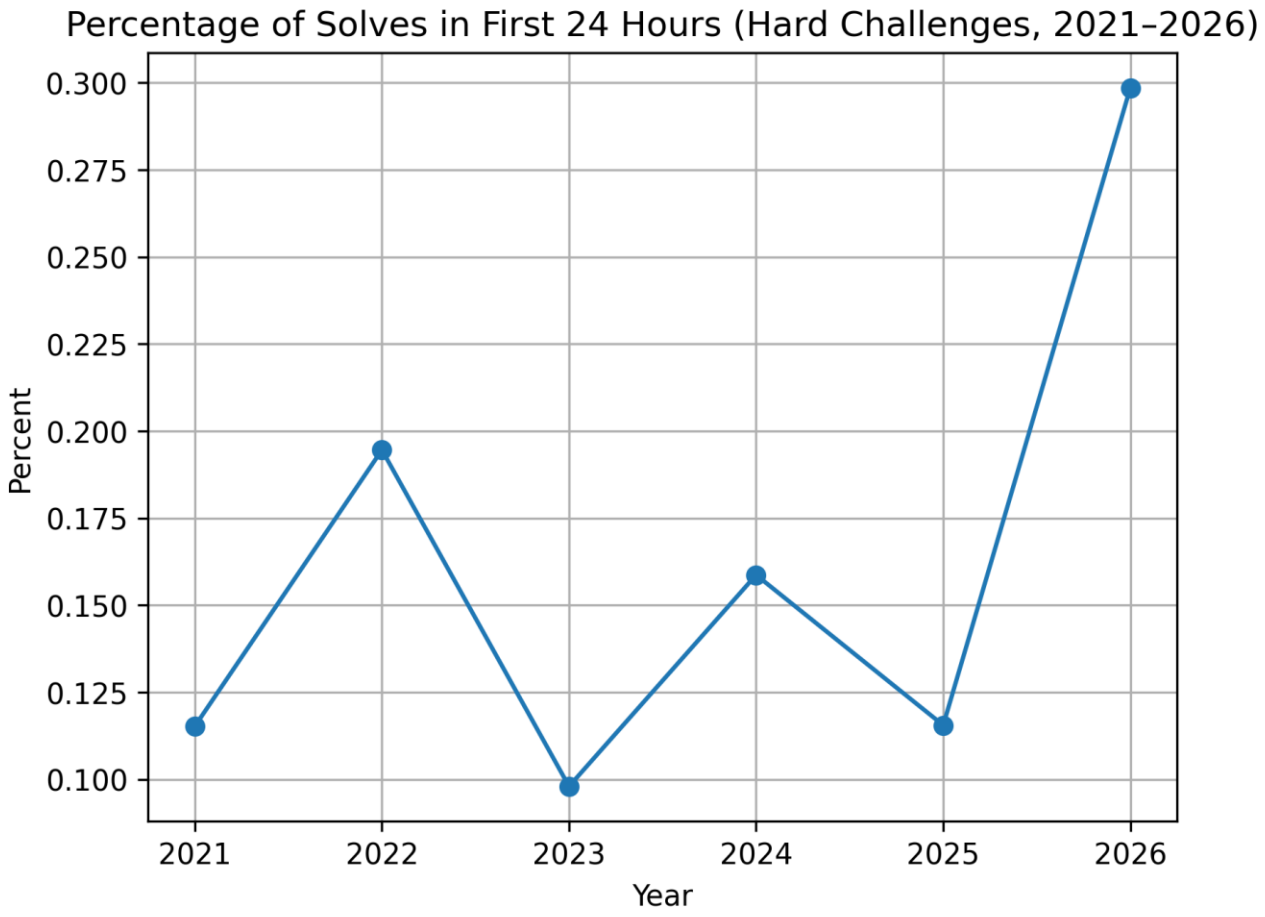

Fig. 4 — Percentage of hard challenge solves occurring within the first 24 hours, by year. The marked increase from 2023 onward indicates a shift from extended, iterative problem-solving to rapid early convergence.

Hard challenges, which had historically shown the most back-weighted solve distributions because they required extended effort, shifted toward the front of the window at a faster rate than easy or medium challenges. A difficulty tier designed to separate competitors who could sustain effort over days was being cleared in hours rather than distributed across days.

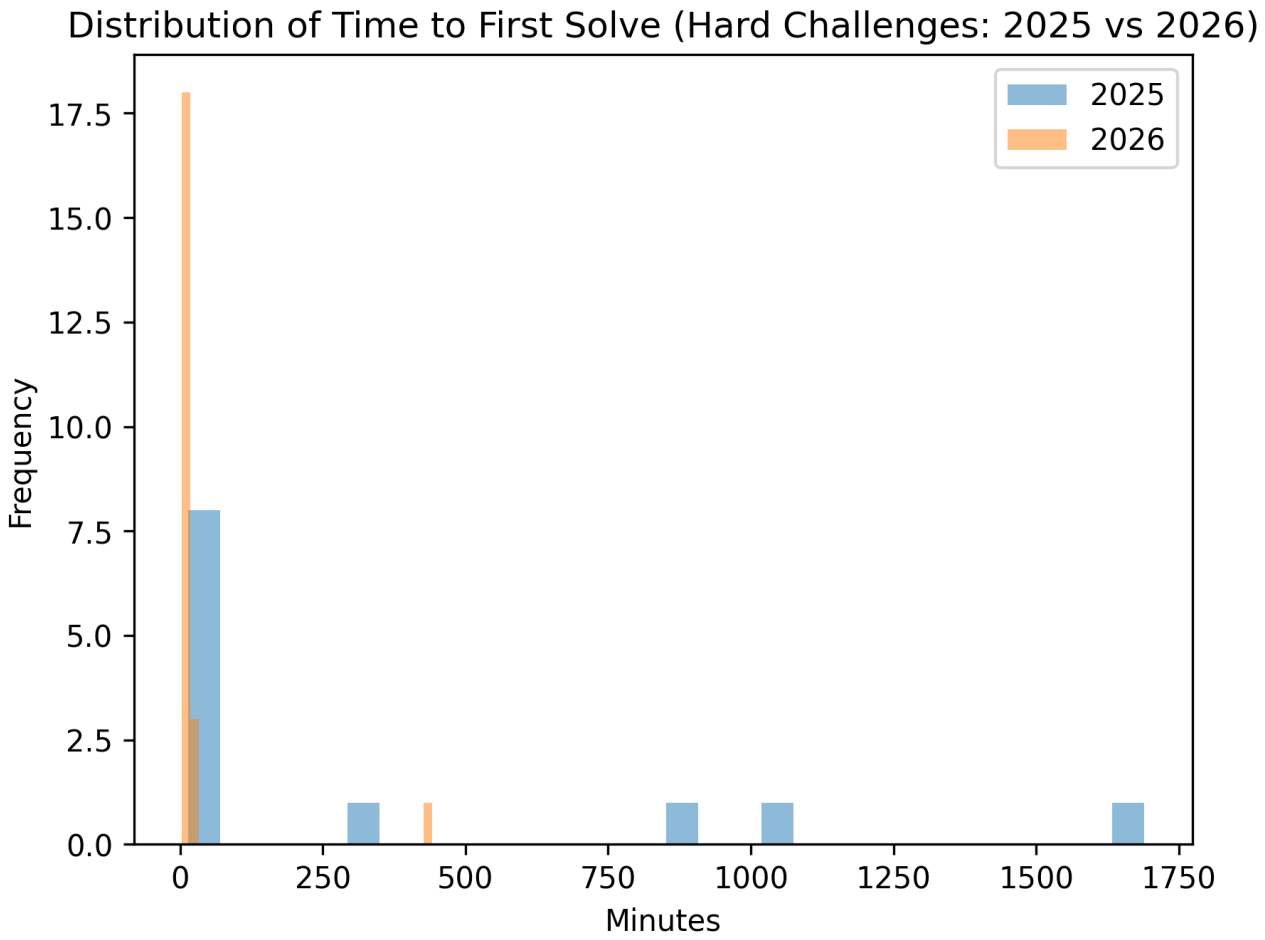

Fig. 5 — Distribution of time to first solve for hard challenges: 2025 vs 2026. The 2026 distribution is heavily concentrated in the first minutes of the competition window, compared to the spread across hours and days seen in 2025.

Speed alone does not establish cause. A faster field could reflect a stronger field. What makes the velocity finding significant is not the speed in isolation but its combination with the ceiling effect in Finding 1: the competitive tier is not just solving faster, it is solving everything, with no variation, in a competition window that is shorter than it used to be.

Taken together with Finding 1, the acceleration in solve timing does not simply indicate a faster field. It indicates a field that is reaching complete solutions earlier, with less variation, and with reduced reliance on extended iterative problem-solving.

Finding 3: The Field Is Changing at Both Ends

Two separate datasets point in the same direction from different angles. New competitors are arriving better prepared than any prior entering class, while returning competitors who came back after AI tools became widely available show a sharp drop in improvement rates relative to those who returned before. Read together, they describe a field changing at both ends at once: a sharply rising entry baseline meeting a flattening improvement curve for those already on the platform. What that

convergence reflects about skill development is the question this finding examines; the data establishes the pattern, not its cause.

Start with who is walking in the door. Each year, the majority of picoCTF participants compete for the first time. Their performance in that first year is a clean signal. The platform has not yet taught them anything. Whatever they bring to the competition, they bring from outside. Tracking how entering cohorts perform over time is one of the most direct windows the data offers into how the broader preparation environment is changing.

For the first four years of this study, that signal was stable. Among participants who engaged actively in their entry year, between 20% and 26% solved at least one hard challenge, and fewer than 8% solved three or more. The exception was 2022, when an expanded beginner track temporarily elevated those figures; setting that year aside, the 2021, 2023, and 2024 entering classes look nearly identical on these measures despite the platform growing substantially across that period. The preparation environment, whatever students were bringing with them, was consistent.

The 2026 entering class is not consistent with that pattern. Nearly half of active first-time participants solved at least one hard challenge. Nearly one in four solved three or more, compared to 6% in 2021. The mean hard challenge solve rate for 2026 new entrants is four times the 2021 figure. These are student accounts that never competed on picoCTF before, so we cannot confirm the platform prepared them for competition.

While this dataset does not directly observe tool usage at the individual solve level, the timing and distribution of solve patterns are consistent with AI-mediated assistance. A student in 2026 who has spent time working through security problems with an AI assistant, asking it to explain concepts, debug approaches, or walk through unfamiliar vulnerability classes, arrives with a different foundation than a student in 2021 who had no such tools available. The entering cohort data cannot confirm what prepared them, but the timing, scale, and abruptness of the shift point toward one plausible candidate: the broad availability of AI tools in the period immediately preceding this cohort's entry.

The returning user data tells a sharper story. Among participants who returned to compete on the same account across multiple years, 84% of those whose second competition fell before AI tools were widely available improved their score percentile. Among those whose second competition fell after 2023, only 41% did. That is a 43-point gap between cohorts separated primarily by when they competed, not by how long they had been on the platform. The pre-AI cohort is n=800, drawn from 2021 to 2022 transitions, and the post-AI cohort is n=3,824, a difference in sample size worth noting when interpreting the comparison. It is highly probable that the larger 2026 cohort includes a broader, potentially less specialized demographic than the 2021 cohort, which would naturally depress aggregate improvement rates. The ceiling effect documented in Finding 1 also explains part of this: when the overall field improves dramatically, maintaining relative standing requires more than marginal improvement. But a 43-point gap is not adequately explained by field-level compression alone. Research on human-AI collaboration has found that tools capable of performing tasks that humans could accomplish themselves often undermine skill development rather than support it, even when deployed with good intentions (Gupta et al., 2025). A returning user who relied heavily on AI assistance in their second competition year may have improved their score without developing the underlying skill that score is supposed to represent. The data cannot confirm that directly. But the gap is large enough that both explanations need to be taken seriously, and both warrant consideration.

Entry baselines are climbing while returning-user improvement flattens, and the gap between them is narrowing faster than any prior year would have predicted. Whether that convergence reflects who is arriving, how they are working, or both is the question Section 3 takes up.

Finding 4: An Emerging Behavioral Mode of Unsupervised Agent Deployment (*preliminary*)

Unlike Findings 1 through 3, which rest on population-scale quantitative records, this finding draws on a narrower and largely qualitative evidence base: submission-pattern logs from a subset of participants, together with self-reported accounts. The full analysis remains in progress, and it is included here as an early signal of a behavioral mode worth tracking rather than a settled result. Among a subset of participants in 2026, the platform logged flag responses at volumes and velocities indicating they were deploying autonomous agents and not reviewing what those agents returned before submitting. A number of participants contacted the CyLab team directly, through email and public Discord posts, to report that their agents had retrieved fabricated strings from external repositories and sites claiming to hold flags for specific challenges and submitted those strings as valid answers without any verification step. While these self-reported cases represent a fraction of the total participant base, they provide a critical qualitative window into how this unmonitored deployment fails in practice. Participants reached out for two reasons: to contest disqualification on the grounds that they had not intentionally cheated, and to request that CyLab pursue removal of the repositories providing false flags. Both responses reveal the same underlying condition. The agent acted as a proxy for the participant but without the participant's understanding of what the proxy had done or where it had retrieved its answer. Research on human-AI collaboration identifies this as one of the most persistent failure modes in AI-assisted work: users calibrate their trust to their initial impression of an agent's competence, and rarely apply the critical scrutiny needed to catch errors that emerge in novel or adversarial conditions (Gupta et al., 2025; Glikson & Woolley, 2020). The agent did not fail to find a flag. It found a false one, retrieved from a source the participant did not know existed, and the participant had no mechanism to know the difference.

Section 3: What This Means for Cybersecurity Education

The findings presented in Sections 1 and 2 describe a measurable shift in performance, behavior, and participant preparation. Before examining the implications of these shifts, the assumption that AI assistance is the primary driver warrants scrutiny. Because the platform does not directly instrument offline tool use, we must evaluate three alternative hypotheses for the observed performance compression: a structural reduction in challenge difficulty, widespread flag leakage, or a sudden influx of highly resourced participants.

These alternatives are difficult to sustain against the platform's six-year record:

- **Consistent Difficulty:** The architectural complexity and difficulty calibration of the 2026 challenge set remained stable. Problems utilized vulnerability classes that historically produced time-to-first-solve metrics measured in days, yet were solved in minutes. It was the velocity of the solutions that changed, not the structural difficulty.

- **Absence of Leak Signatures:** Historically, leaked solutions produce clustered, delayed solve spikes among lower-tier participants. In 2026, the velocity shift occurred immediately upon competition opening, on novel challenges, resulting in score compression at the absolute top of the field. This indicates a real-time capability to solve new problems, rather than the retrieval of static flags.
- **Demographic Divergence:** If 2026 simply attracted a more talented demographic, performance would rise uniformly. Instead, we observed a stark divergence: entering cohorts arrived with a sharply raised baseline, while the relative improvement rates of returning competitors plummeted.

Taken together, these alternatives fail to explain the magnitude, timing, and structure of the 2026 data. The observations align most closely with the broad availability of advanced reasoning models, pointing to a field that has fundamentally changed its approach to problem-solving.

What the data must now address is the central question these shifts raise: **what is a cybersecurity competition measuring in the presence of capable AI systems, and what should it measure going forward?**

For much of its history, competitive cybersecurity education has operated under a clear assumption. Performance in a Capture the Flag competition reflects a combination of knowledge, persistence, and the ability to apply tools effectively under time constraints. Over time, the role of tooling has expanded. Scriptable frameworks, automated scanners, and public exploit repositories have all reduced the amount of manual effort required to solve certain classes of problems. However, these tools historically operated within bounded domains. They accelerated specific steps in the problem-solving process, but they did not replace the need for human direction. They were instruments, not substitutes.

The data from 2026 suggests that this distinction is no longer stable.

AI systems, particularly those capable of multi-step reasoning and tool use, appear to have reached a level where end-to-end problem solving is achievable across many CTF challenge types. External research confirms this is not a platform-specific observation. By late 2025, only months after the April Hack The Box event, that trajectory had advanced sharply. Across five major CTF competitions through the rest of the year, the autonomous agent evaluated by Mayoral-Vilches et al. (2025) reached peak rankings of #1 at multiple events and outpaced the large majority of human teams. This advantage was driven primarily by early-window velocity, reaching key point thresholds at one event some 37% faster than the top-five human average, with per-event solve rates as high as 91%. The edge was concentrated in the opening hours of competition rather than sustained to the final standings. At the top of the competitive field, timing has already displaced score as the primary differentiator. The same pattern appeared in our own data a year later. In this environment, the marginal advantage gained from mastering an additional tool is qualitatively different from the advantage gained by deploying an autonomous or semi-autonomous system that can explore, iterate, and produce candidate solutions at machine speed. What was previously an acceleration of human effort is increasingly a partial replacement of it.

This does not render cybersecurity competitions obsolete. It does, however, change what their outcomes signify.

A first-place finish in a global CTF has historically been interpreted as a proxy for individual or team skill at the highest level. Under current conditions, that interpretation becomes ambiguous. When multiple

teams achieve full completion and rankings are determined primarily by speed, the competition is no longer differentiated based on depth of understanding. Instead, it measures efficiency of execution, which may include factors external to human cognition: automation pipelines, AI integration, and infrastructure optimization. These are not illegitimate skills, but they are different from the competencies that cybersecurity education has traditionally aimed to cultivate. As Mayoral-Vilches et al. (2025) argue, Jeopardy-style CTFs have become computational exercises rather than genuine security skill assessments, with timing becoming the primary differentiator once AI agents achieve near-ceiling solve rates.

These observations do not, on their own, prescribe a response; they establish a measurement problem. What follows is the response that problem points toward: reasoning from the data, not a conclusion the data settles.

This creates a tension between two goals that have long coexisted within CTF design: competition as measurement and competition as pedagogy.

The competitive format remains compelling. It motivates engagement, provides clear goals, and creates a shared experience that draws participants into sustained practice. At the same time, the conditions under which competition outcomes can be interpreted as indicators of individual understanding are eroding in large-scale distributed competitions where participants are unknown to each other and where the platform can observe behavior but cannot verify the identity or agency behind it. The platform data shows that we cannot reliably distinguish between participants who understand a solution and those who can obtain one. Attempts to restore that distinction through enforcement, by detecting or prohibiting AI use, may be difficult to sustain at scale. The behaviors described in Section 1 already demonstrate that participants are experimenting with systems they do not fully control. As AI capabilities continue to improve and become more accessible, that distinction will become increasingly difficult to maintain.

This is not an unfamiliar problem in other fields where credentials carry high stakes (e.g. medicine, aviation, and law), practitioners have long had to solve a specific design challenge: how to confirm that a person has a capability independent of the conditions under which they demonstrated it. Those fields developed assessment conditions that required human reasoning directly and treated the results of those assessments as the authoritative signal of what a person could do. A credential earned under those conditions carries different meaning than a performance score produced under undetermined ones. Cybersecurity has developed significant certification infrastructure. What the data in this paper makes visible is that the gap between assisted performance and verified human capability has widened quickly enough to become a practical problem. A learner whose competition score cannot be distinguished from what an AI agent would produce can potentially be distinguished by how they perform under conditions that require explanation, demonstration, and independent reasoning. Building and credentialing those conditions at scale is the work this data points toward. Rather than attempting to preserve a prior state, the case for cybersecurity education adapting to the conditions that now exist is correspondingly strong.

One implication is a shift in where meaningful competition occurs. At a global level, where participants are effectively anonymous and unobservable, competition will continue to trend toward automation-assisted performance. In co-located or socially bounded environments, different dynamics apply. Classrooms, clubs, camps, and local events share context. They can observe one another's work, discuss approaches, and establish norms around tool usage, including AI. In such environments, accountability is social rather than technical. Scholars advocating for the transition to Attack and Defense formats

make a related point: dynamic adversarial elements that require real-time service defense, adaptive patch management, and strategic decision-making under pressure introduce capabilities that resist simple automation in ways that static Jeopardy-style challenges no longer do (Mayoral-Vilches et al., 2025; Balassone et al., 2025).

This suggests a structural evolution: global competitions function as broad access learning platforms, while meaningful competitive comparison shifts toward local cohorts where norms can be established and enforced through presence rather than instrumentation. A global scoreboard reflects aggregate performance under mixed conditions. A classroom scoreboard reflects a more controlled learning environment. The same challenge set can support both modes, but the interpretation of results differs.

This reframing connects directly to the second implication of the data: the widening gap between the ability to use AI systems and the ability to understand and appropriately supervise them.

That gap is not new to researchers studying human agency in AI-mediated environments. Xie and Cullen (2026) distinguish between AI literacy, which improves understanding of how AI systems function, and what they term well-being efficacy, the integrated capacity to preserve agency, coherence, and ethical clarity when operating within systems that exceed direct human comprehension. Their argument is that knowledge of how to use AI is necessary but insufficient. What the observed behaviors in this data make visible is the same insufficiency in a cybersecurity-specific context: participants are deploying systems whose outputs they cannot fully evaluate, predict, or constrain. In a domain where incorrect assumptions can produce unintended system interactions, that is not an abstract concern.

Cybersecurity education has always required a grounding in how systems behave under adversarial conditions. That requirement now extends to AI systems themselves. Students must not only understand networks, binaries, and protocols, but also the limitations, failure modes, and unpredictability of the tools they are using to interact with those systems. The problem is no longer simply “how do I exploit this vulnerability,” but also “how do I evaluate whether the system proposing this exploit is correct, safe, or applicable in this context.”

This represents a convergence between cybersecurity literacy and AI literacy, and more specifically between cybersecurity education and the kind of reflective, supervisory capacity that Xie and Cullen situate within a well-being framework.

Educational programs that treat AI as an external tool risk reinforcing superficial competence: the ability to obtain answers without understanding them. Programs that instead incorporate AI into the learning process, examining its outputs, identifying its errors, and requiring explanation, can leverage its strengths while mitigating its risks. The goal is not to prevent students from using AI, but to ensure that its use deepens rather than replaces understanding.

That goal has implications for assessment design. If the objective is to measure understanding, then assessment must require demonstrations of understanding. In practice, this may include requiring written or oral explanations of solutions, analyzing incorrect or incomplete outputs, or structuring challenges such that the first plausible answer is not sufficient without interpretation. These approaches are not new, but their importance increases sharply in an environment where correct answers can be obtained without comprehension.

It also has implications for how foundational concepts are taught. The rapid development of large language models has outpaced the ability of many learners to form a coherent mental model of how these systems function. While full mechanistic understanding remains an active area of research, there

is value in teaching what is known in a deliberate and accessible way. Concepts such as probabilistic generation, training data influence, and common failure modes can be introduced alongside traditional cybersecurity topics. This does not require slowing innovation broadly, but it does require creating educational spaces where understanding is prioritized over speed.

The pace of development in AI systems has been extraordinary and that pace is unlikely to slow. The question for cybersecurity education is not whether to keep up, but how to respond in a way that produces practitioners who can operate responsibly within an environment defined by capable AI. CTFs will continue to play a role in that response. Their value may shift from serving primarily as global competitive benchmarks to serving as structured environments for exploration, practice, and guided learning. The competitive element will remain, but its locus will change. The educational function, which has always been present, becomes primary.

In that sense, the emergence of AI-assisted solving does not diminish the importance of cybersecurity education but clarifies it.

Section 4: How CyLab Is Responding

THE SHIFT FROM COMPETITION TO EDUCATION PLATFORM PRECEDED THE AI INFLECTION POINT

The responses described in this section reflect CyLab's current institutional direction and the platform and content development work underway in response to the findings presented above. This section remains under active development as the findings in this paper are reviewed and extended and will be updated in a subsequent version. The responses described in this section are not reactive adjustments to the 2026 data. They reflect an institutional orientation toward education over competition that CyLab began acting on in 2020, and which the findings in this paper serve to validate and accelerate. Understanding that orientation requires a brief account of what the platform was before it became what it is now.

From 2013–2017, picoCTF was structured as a discrete annual event. The competition window ran for ten to fourteen days, and when it closed, participant access closed with it. Educators who wanted to build coursework around the challenge content had no persistent resource to reference, and participants who wanted to continue developing their skills after the competition ended had no mechanism to do so within the platform itself. The implicit assumption embedded in this model was that competition was the primary use case and that educational value was a byproduct of competitive participation rather than a goal.

That assumption began to show its limits in 2018 and 2019, when the competition platform was reopened on a limited basis following competition close to allow continued practice access. Rather than tapering off after the competitive event, participation persisted. The 2019 competition itself drew 39,349 participants during its active window, a figure consistent with the steady growth the event had shown since its founding. By the time the 2019 platform was closed in late 2020, however, internal records showed 115,879 registered users, of whom 27,151 had formed 12,542 teams, representing nearly three times the competition participant count generated by making the content persistently accessible rather than time-gating it behind the competitive event. More tellingly for the educational trajectory, 2,227 classrooms had been established by 2,685 teachers, a figure that could not have accumulated within the

competition window and reflected sustained engagement with the platform as a teaching resource across the intervening months. The evidence pointed to a straightforward conclusion: significantly more participants wanted the content than wanted the competition, and the annual open-and-close model had been constraining that demand artificially.

The 2019 challenge architecture itself offered a further signal. The dependency graph underlying that competition contained 122 challenges organized around 11 entry points, with prerequisite chains extending up to nine levels deep and 13 challenges requiring completion of multiple prerequisites before they became accessible. That structure describes a curriculum, one in which learners progress through sequenced competency development rather than selecting freely from a pool of parallel problems. The two most advanced challenges in the graph each required nine sequential steps to reach, a design decision that only makes sense if the intent is to develop understanding incrementally rather than to present a maximally open competitive environment.

In 2020, CyLab transitioned the platform to continuous availability, retaining the annual competition as an anchor event while making the full challenge library accessible year-round. The findings presented in this paper provide a more urgent rationale for that decision than was available at the time it was made. If competition outcomes are becoming increasingly difficult to interpret as reliable measures of individual skill and understanding, the sustained educational function of the platform carries more weight, not less. A participant who uses AI assistance to clear a hard challenge in 2026 without developing an understanding of the underlying vulnerability has passed through the competition without gaining the competency it was designed to develop. A participant who uses that same challenge as a starting point for structured inquiry, examining what the tool did, testing where it fails, and building a working model of the underlying system, has used the platform for its intended purpose regardless of how or whether they appear on a scoreboard.

RESPONSES CONNECTED TO FINDINGS

CyLab's response to the findings in this paper is organized around four areas of platform and content development, each connected to a named finding from Section 2. These responses are at different stages of implementation, and where work remains in progress, that is noted.

Finding 1	Competitive differentiation at the top of the field has collapsed. Score no longer functions as a discriminating metric among the top 50 teams.

In response to the erosion of competitive differentiation documented in Finding 1, CyLab is developing a structured badge and milestone system tied to learning path progression across the Academy's tracks in Cybersecurity, CTF, AI Security, Blockchain Security, and Economics of Security. This system is a work in progress, and the central design challenge it is navigating is how to distinguish meaningfully between recognition that rewards sustained engagement and markers that reflect demonstrated competency at a level worth signaling to others.

The tracks themselves reflect where the field is heading. Cybersecurity and CTF represent the core of what competitive security education has always developed. AI Security, as described above, addresses

the emerging requirement that practitioners understand and supervise the systems they deploy. The two newest additions point in a direction the Academy's six years of learner data suggested before the content existed to address it.

Blockchain Security sits at the intersection of cryptographic systems and economic incentive design. The challenges in this track are not primarily about finding vulnerabilities in code. They are about understanding systems where the adversary is not breaking the rules but playing them correctly in ways the designer did not anticipate. Smart contract exploits, maximal extractable value, and protocol attacks that weaponize rational self-interest rather than implementation flaws require a different cognitive frame: not what did the developer get wrong, but what did the designer fail to predict about how a fully rational actor would behave. Four challenges launched in 2026 across two difficulty tiers. The foundational medium challenges, Smart Overflow and Access Control, recorded 1,144 and 1,344 solves respectively with thumbs up ratings of 96% and 94%, indicating that participants found the material accessible and engaging in a subject area with no prior presence on the platform. The hard challenges, Reentrance and Front Running, recorded 950 and 1,022 solves with platform-wide thumbs up ratings of 95% and 93%. Ratings at this level on hard-tier challenges are uncommon across the platform and reflect a self-selected population of learners who arrived with relevant prior knowledge and strong motivation to engage with the material. Combined, the four challenges recorded 4,460 solves in their debut year. Discord commentary from the community indicated immediate demand for more.

Economics of Security extends that frame further. It asks not just how rational actors exploit systems, but how security systems can be designed to remain robust when interacted with by adversaries whose only constraint is self-interest. This is the same question that underlies AI alignment research, mechanism design, and adversarial machine learning. The faculty foundation for this track is complete, and implementation is planned within the next eight to twelve months. The connection to the AI Security track is intentional: a practitioner who understands both how economic incentives shape adversarial behavior and how AI systems optimize for reward functions under constraints is equipped for a category of security work that neither track addresses alone.

That distinction is not merely aspirational. It already exists in the challenge architecture the platform has built over six years. Hard challenges on the platform represent competency thresholds rather than difficulty gradations applied to manage participation flow. A challenge like Solfire, which requires expert-level command of blockchain security vulnerabilities to solve, is not approachable through persistence or procedural pattern-matching alone, and completing it reflects knowledge that would be recognized as meaningful by practitioners in the field. Solfire was released in 2022, before AI tools became capable of approximating expert-level security reasoning, and the knowledge threshold it represents reflects practitioner judgment rather than a design response to automation. The goal of the badge and milestone system is to make that kind of meaning legible at scale, so that a learner who progresses from the AI Security foundations path through to an agentic security track carries a record that accurately represents the conceptual territory they have covered: security behaviors, vulnerability classes, failure mode analysis, and the systems-level reasoning required to build more resilient AI-integrated systems. The challenge infrastructure that would underpin such a system already exists; the work now is in developing the assessment and recognition layer that makes it coherent as a learning credential. What makes this system meaningful in the current environment is not simply that it recognizes achievement. The challenge architecture underlying it was built before AI tools became capable of approximating end-to-end problem solving, and the platform has six years of behavioral data on how human learners actually progress through that architecture. A credential anchored to demonstrated progression through that record is not a record of what a learner obtained in a single session. It is evidence of a sustained pattern of engagement with the material. The distinction the badge system is designed to

capture is not speed of completion. It is depth of coverage across a sequenced body of knowledge that builds on itself. That distinction matters beyond the platform. At the competitive tier, where the data shows that scores no longer differentiate between teams, a performance record built on verified human capability carries a signal that a leaderboard rank no longer does. For employers, for educational institutions, and for the cybersecurity community evaluating who is ready for adversarial real-world work, the question is not whether someone can produce the right answer. It is whether they can produce it when the tools are not there, or when the tools are wrong.

Finding 2	Solve velocity has accelerated sharply. The competition window is being completed in hours rather than days at the top of the field.

The velocity findings in Finding 2 raise a design question that extends beyond integrity enforcement: if the global CTF competition increasingly measures automation efficiency rather than problem-solving depth, what formats can restore the interpretive value of competitive performance as an educational signal? CyLab is developing complementary formats that place weight on demonstrated understanding alongside speed, including challenge structures where the first plausible solution is insufficient without an accompanying explanation of the method and its limits, and locally administered cohort events, like Classroom CTF, in which social accountability around tool norms replaces the technical instrumentation that is unlikely to scale in a distributed anonymous environment. The global scoreboard is not being retired; its role as a broad-access entry point and motivation mechanism remains intact. What is changing is the weight placed on global competitive rank as an indicator of individual learning outcomes, relative to these more observable local formats.

Finding 3	Entering cohorts are arriving better prepared while returning users are improving at lower rates relative to the field. The floor is rising faster than individuals are climbing.

The divergence between entering cohort performance and returning user improvement documented in Finding 3 points to a structural gap that content design can address. CyLab is investing in challenge content and structured learning paths that sit at the intersection of AI-assisted performance and the independent understanding that should follow from it, with particular attention to AI security content that engages participants with how these systems function, where their outputs are unreliable, and how to evaluate them critically rather than accepting them as authoritative. The objective is not to restrict tool use but to ensure that the platform creates conditions in which tool use can become a starting point for deeper engagement rather than a substitute for it. Whether that outcome is achieved at scale is an empirical question. The platform's behavioral record is the infrastructure for beginning to answer it.

Finding 4	Participants are deploying autonomous agents in competition environments, in some cases without full understanding of what those systems are doing.

The agent deployment behaviors documented in Finding 4 present an opportunity as well as an enforcement challenge. Participants who deploy autonomous systems in competitive environments without fully understanding what those systems are doing represent a gap that CyLab is positioned to address through content development rather than solely through disqualification. The gap is not primarily one of intent. It is one of technical preparation. CyLab is building an AI Security track that begins with foundational concepts and is designed to progress toward the technical capacity to specify, test, and audit agentic systems. The agent behavior component represents the direction that track is headed, informed by the behavioral evidence in this paper. The same discipline that security-focused curricula apply to code, defining expected behavior, testing against adversarial conditions, identifying failure modes, and verifying outputs, applies equally to agentic systems. Students on this track do not simply observe what their agent did. They learn to specify what it should do, constrain what it is permitted to do, interpret what its outputs mean, and detect when it has failed in ways that are not immediately visible. This is the same reasoning process a developer applies when writing secure code. The difference is that the artifact being tested is an AI system rather than a function or a module. This framing reflects a broader position the field is beginning to recognize: responsible AI is not a governance layer applied after development. It is a technical engineering discipline that belongs in the development process from the start, the same way security now belongs in software engineering by design. A developer who ships code without considering what an attacker will do with it is technically incomplete, not just negligent. A developer who deploys an agentic system without the capacity to specify, constrain, and audit its behavior is in the same position. The technical standards for what responsible AI means in practice are still being established, which is precisely why practitioners who understand both security and AI systems need to be part of building them. CyLab's AI Security track is designed to develop that capacity at the level where it is most needed and currently most absent: early in a practitioner's formation, before the habits of unexamined AI use are established. The specific structure of this track will be refined as the behavioral analysis underlying Finding 4 continues and a fuller taxonomy of these deployment patterns is developed.

THE BROADER CONTEXT

The gap this paper documents sits at the intersection of three skills that education and workforce development have not yet treated as a coherent set. The first is technical skills, the foundation cybersecurity education has always built. The second is judgment: the capacity to evaluate an AI output critically, to recognize when it is wrong, incomplete, or inapplicable, and to decide how to act on it. The third is responsible AI as a technical engineering discipline. Not a policy layer and not an ethics checklist, but the practical capacity to specify agent behavior, set meaningful constraints, interpret outputs, detect failures, and maintain human oversight of agentic systems in operational contexts. That third skill is where the field currently has the largest gap, and the analogy that clarifies why is already familiar to anyone in software engineering. Security by design, the principle that security cannot be added to software after the fact and must be built in from the start, took decades to become standard practice and met significant resistance along the way. A developer who does not consider security during development is not just negligent; they are technically incomplete. The field eventually encoded that judgment in practice, in hiring, and in curriculum. The same transition is now required for AI. A developer who builds or deploys agentic systems without the capacity to specify, constrain, and audit their behavior may hold the right values and still cause the wrong outcomes. Ethical intent does not substitute for technical capability, and technical capability without ethical grounding is capable of its own category of harm. The field needs practitioners who bring both, and the data in this paper documents what happens when they do not. The 2026 competition data makes that gap visible in a

specific and measurable way: participants deployed systems they could not supervise, producing outputs they submitted without understanding. Current evidence suggests that AI-related workforce development has remained oriented toward deployment rather than oversight (OECD, 2025). The consequences are more acute in cybersecurity than in most fields because the systems under analysis are adversarial by design, and because overreliance on AI outputs without sufficient critical evaluation has already been identified as a direct operational risk (ISACA, 2024; Achuthan et al., 2024). Research on AI-assisted learning has further established that improved task performance under AI assistance does not reliably translate into the underlying competency required to function when that assistance is absent or incorrect (Bastani et al., 2025). CyLab's platform, with six years of behavioral data on how learners engage with security problems under competitive conditions, is positioned to generate the empirical foundation that content addressing all three of these skills will require. What kinds of challenges develop judgment rather than just technique? At what point in a learner's progression does responsible AI practice need to be introduced? What behavioral signals distinguish a student who understands an AI-assisted solution from one who obtained it without understanding? The findings in this paper establish that these questions are answerable, and that the behavioral record needed to answer them exists at scale. That is the foundation the next stage of this work builds on.

Addressing that gap at scale requires an organization with longitudinal behavioral data on how learners actually engage with AI-assisted problem-solving, a platform capable of delivering structured content across a population large enough to generate statistically meaningful signal, and an institutional standing that makes its outputs credible to the research, policy, and practitioner communities that need to act on them. CyLab's platform, with more than one million registered learners, over sixteen thousand teachers, fourteen thousand active classrooms, and more than ten million challenge solves accumulated across six years of continuous operation, occupies that position. The data that produced the findings in this paper was generated by that platform. The infrastructure required to respond to those findings in a way that reaches the learners who need it most is the same infrastructure.

The transition from annual competition to perpetual education platform was not designed as a response to AI-mediated learning. It was a response to the evidence available in 2018 and 2019 that the platform's educational value was constrained by its competitive framing, and that participants and educators were asking for something more durable. The 2026 data clarify, with considerably more urgency, why that reorientation was the right one. The competitive event will continue to serve as an annual benchmark and entry point. The educational platform that surrounds it is where the more important work is now happening.

References

Achuthan, K., Ramanathan, R., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: Current trends and future research directions. Frontiers in Big Data. https://doi.org/10.3389/fdata.2024.1497535

Balassone, F., Mayoral-Vilches, V., Rass, S., Pinzger, M., Perrone, G., Romano, S. P., & Schartner, P. (2025). Cybersecurity AI: Evaluating agentic cybersecurity in attack/defense CTFs. arXiv. https://arxiv.org/abs/2510.17521

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, O., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gupta, P., Nguyen, T. N., Gonzalez, C., & Woolley, A. W. (2025). Fostering collective intelligence in human–AI collaboration: Laying the groundwork for COHUMAIN. Topics in Cognitive Science, 17(2), 189–216. https://doi.org/10.1111/tops.12679

Hack The Box. (2025). AI vs human CTF competition results. https://www.hackthebox.com/blog/ai-vs-human-ctf-2025

ISACA. (2024). AI and automation in cybersecurity: Future skilling for efficient defense. ISACA Journal, Volume 3. https://www.isaca.org/resources/isaca-journal/issues/2024/volume-3/ai-and-automation-in-cybersecurity-future-skilling-for-efficient-defense

Mayoral-Vilches, V., Navarrete-Lozano, L. J., Balassone, F., Sanz-Gómez, M., Veas Chavez, C. R. J., del Mundo de Torres, M., & Turiel, V. (2025). Cybersecurity AI: The world's top AI agent for security capture-the-flag (CTF). arXiv. https://arxiv.org/abs/2512.02654

OECD. (2025). Bridging the AI skills gap: Is training keeping up? OECD Publishing. https://doi.org/10.1787/66d0702e-en

World Economic Forum. (2025). Future of Jobs Report 2025. https://www.weforum.org/publications/the-future-of-jobs-report-2025/

Xie, Y., & Cullen, W. (2026). Beyond procedural compliance: Human oversight as a dimension of well-being efficacy in AI governance. Proceedings of the Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26).