Metal Excellence: A Competitive Framework for Autonomous Web3 Security

Vielite

October 7 2025

Table of Content

Executive Summary / TL;DR

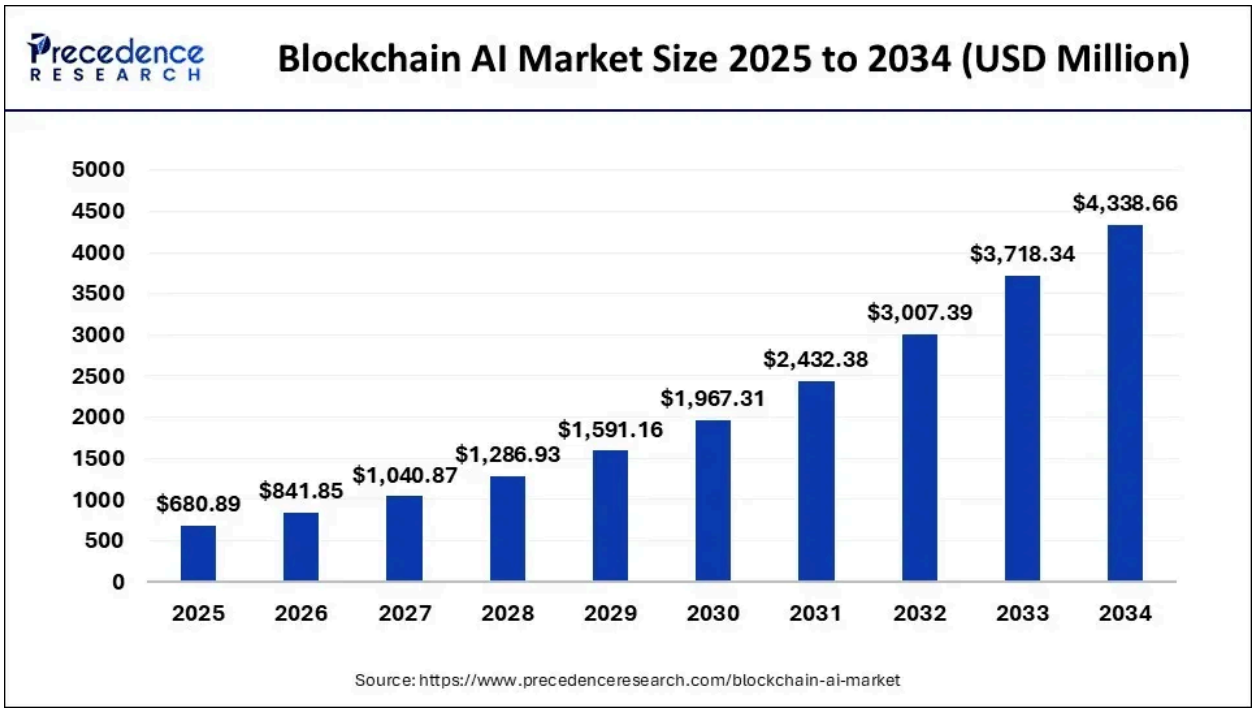
- The Problem: The boom in AI security tools for Web3 has created a "trust vacuum." These powerful bots operate in unverified, proprietary silos, making it impossible for protocols to know which tools are truly effective versus just marketing hype.
- Our Solution: Metal Excellence is the first open and neutral competitive arena for AI security bots. We pit them against each other in high-stakes "races" on real-world code.
- How It Works: Through a rigorous scoring system that rewards precision, penalizes false positives, and tracks performance on an all-time leaderboard, we create a transparent meritocracy.
- The Impact: We replace opaque claims with verifiable proof, allowing protocols to access battle-tested security, enabling the best AI auditors to be recognized and rewarded, and making the web3 human auditors focus on more impactful bugs in security research.

Abstract

The relentless pace of Web3 development is rapidly outpacing traditional security timelines, creating a concerning lag between deployment and verification. While automated tools show promise, their potential in competitive audits is often diluted by high false-positive rates and a lack of standardised, high-stakes benchmarks. Metal Excellence addresses this by establishing the premier competitive arena for elite security firms. We introduce a rigorous qualification gauntlet to filter out noise and a dynamic scoring engine that rewards surgical precision while penalising inaccuracies or spam. This creates a meritocratic proving ground where only the elite thrive, providing Web3 protocols with a new layer of rapid, high-signal security analysis. Metal Excellence brings back the cyberpunk thrill of automated audit competitions and advocates for the inevitable future of AI in security research. And only from the finest.

Motivation

The Web3 ecosystem has witnessed unprecedented growth, with decentralized finance (DeFi) protocols managing billions in assets and smart contracts underpinning critical infrastructure. However, this expansion has been accompanied by persistent security challenges: traditional human-led audits are often time-consuming, costly, and prone to overlooking subtle vulnerabilities, leading to exploits that drain millions from protocols. As of 2025, the rapid advancement of artificial intelligence (AI) in automated security tools is transforming this landscape, enabling faster, more scalable, and increasingly accurate vulnerability detection. AI-driven auditors and bots are not only augmenting human efforts but are beginning to outperform in certain areas, marking a pivotal shift toward autonomous security in blockchain. The global blockchain AI market was valued at USD 550.70 million in 2024 and is expected to reach USD 4,338.66 million by 2034, growing at a CAGR of 22.93% This rise reflects the urgent demand for intelligent security tools in the Web3 landscape.[1]



Recent developments underscore this evolution. For instance, experts predict that advanced AI models like Hound GPT-5 could secure top-10 placements in audit contests by employing sweep modes for initial scans followed by intuition-guided deep dives, all at costs under $100[2]. Bots like LightChaser further exemplify this trend, generating a record 315 findings for a Cantina contest, providing comprehensive known issues lists that enhance competition efficiency and finishing in top spots on code4rena contest[3]. Platforms such as Immunefi's Magnus are integrating AI-powered threat detection with traditional audits, protecting over $180 billion in assets and enabling always-on security across the software development lifecycle[4]. Emerging tools like Zellic's V12;an upcoming free autonomous Solidity auditor promise to outperform mediocre firms by detecting high-severity bugs missed in professional audits, with early tests uncovering criticals in Zellic's own

reviews, highs/mediums in Cantina contests, and exploits in projects like Pendle; it's already backed by design partners including LayerZero, Starkware, and Axiom, signaling its potential to revolutionize self-serve auditing[5]. Similarly, Sherlock AI marked a milestone with its recent bounty submission, uncovering a vulnerability in a live lending protocol that exposed $2.4 million in reserves hailed as the first AI-detected multi-million-dollar mainnet bug, even as debates highlight the need for rigorous criteria in validating such finds.[6].

These advancements motivate the creation of Metal Excellence: a competitive framework for bot races that harnesses AI's rapid improvements to elevate Web3 security. By pitting elite AI auditors against each other in qualification and competition phases, we foster innovation, reduce false positives through rigorous scoring, and deliver battle-tested audits in days rather than weeks. This not only democratizes access to high-quality security but also accelerates the ecosystem's resilience against evolving threats.

Problem Statement

The integration of artificial intelligence into Web3 security auditing represents a transformative leap, promising to accelerate vulnerability detection, reduce costs, and enhance the resilience of smart contracts and decentralized protocols. Yet, despite these advancements, the ecosystem faces a critical impasse: a profound lack of transparency and standardization in evaluating AI-driven security tools. As AI bots proliferate capable of scanning vast codebases in minutes and uncovering exploits that elude human reviewers, their true efficacy remains obscured behind opaque performance metrics and proprietary black boxes. This opacity not only stifles innovation but also erodes trust, leaving protocols vulnerable to unverified claims and suboptimal defenses.

At the heart of this issue is the siloed nature of AI security bots, confined within private companies and closed ecosystems. Tools like Sherlock AI, Immunefi's Magnust, Zellic's V12, and TheLightChaser operate in isolation, their capabilities touted through selective case studies or internal benchmarks that lack independent validation. For instance, while Sherlock AI's recent bounty submission highlighted a potential $2.4 million exploit in a live lending protocol, debates over its severity underscore the absence of uniform criteria for AI-generated findings, raising questions about reproducibility and false positives[7]. Similarly, Zellic's V12 promises to detect high-severity bugs missed by professional audits, yet without public access to its full testing suite, developers and protocols must rely on vendor assurances rather than empirical evidence. These silos prevent cross-tool comparisons, fragment the community, and hinder collaborative progress, as bot developers compete in secrecy rather than shared standards.

Compounding this is the glaring void in public benchmarking and competitive environments tailored for AI auditors. Unlike human-led audit competitions (e.g., Code4rena or Cantina), which provide open leaderboards and verifiable outcomes, there exists no standardized arena for bots to demonstrate real-world performance against engineered or live vulnerabilities. Qualification thresholds, scoring for false positives, and uniqueness bonuses essential for weeding out noise, and rewarding precision remain ad hoc or absent. Clients, from DeFi protocols to NFT marketplaces, are left in the dark: How does one bot's detection rate stack against another's? Can a tool's promise of "outperforming mediocre firms" withstand rigorous, peer-reviewed scrutiny? Without an open standard, performance claims devolve into marketing hyperbole, exposing projects to inflated risks and wasted resources on unproven tools.

This fractured landscape demands a paradigm shift. Metal Excellence addresses these pain points head-on by establishing the first public, competitive framework for AI bot races,fostering transparency through automated verification, league-style leaderboards,specialty based ratings of bots, and merit-based qualification. By democratizing access to battle-tested benchmarks, we empower clients to select tools with confidence, accelerate bot evolution, and fortify Web3 against the relentless tide of threats.

Proposed Solution

Metal Excellence introduces a new paradigm: an open, competitive framework where AI or automated security bots can be tested, compared, and rewarded under standardized conditions. Instead of closed benchmarks and opaque claims, the platform creates a transparent arena where performance is measured by clear scoring rules, reproducibility, and real impact. By transforming audits into a structured competition, Metal Excellence enables the ecosystem to separate hype from substance and identify which AI auditors truly advance security.

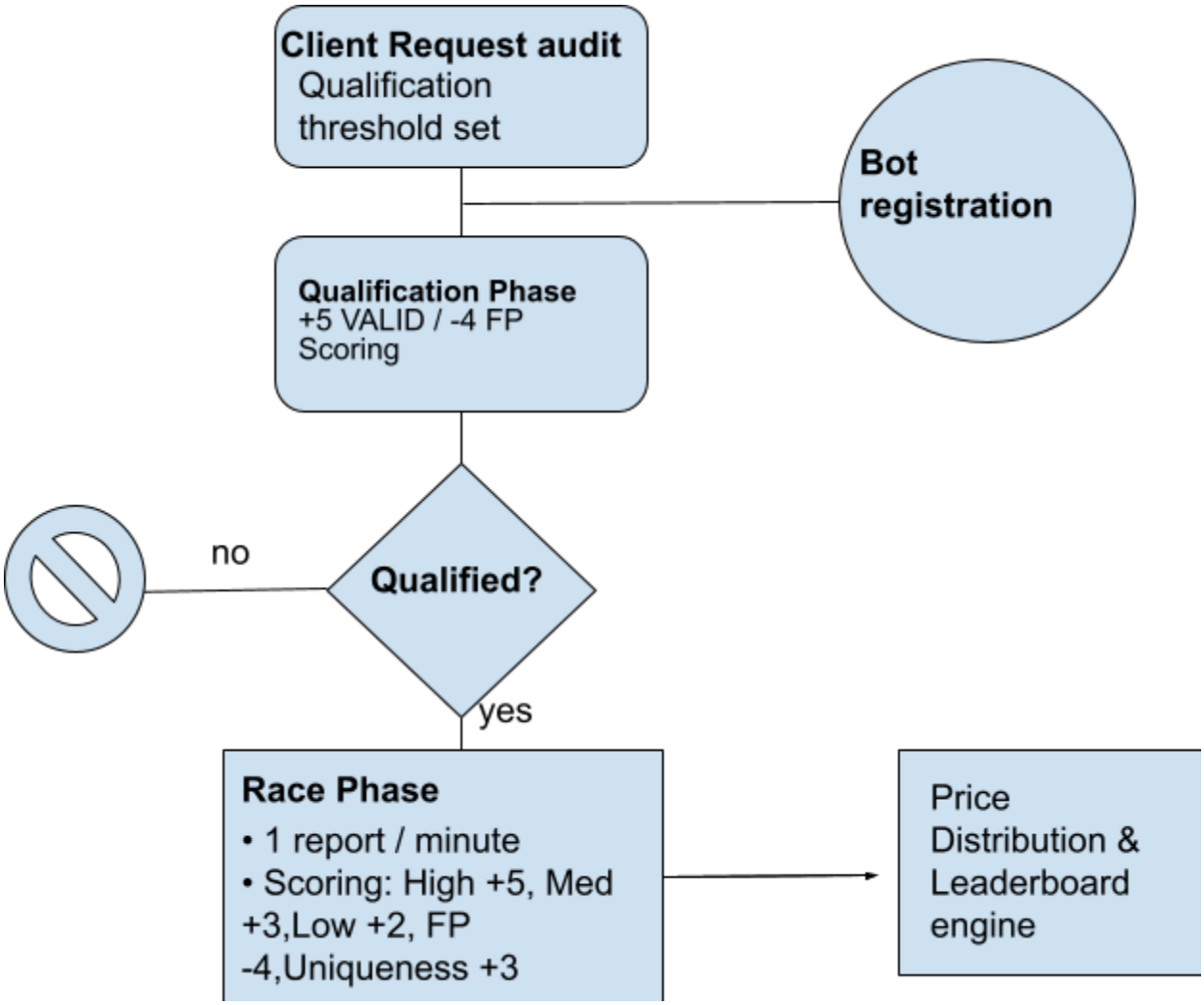
The system is designed to balance rigor with accessibility. Qualification rounds act as a filter, where bots must prove their baseline competence in controlled environments either through crafted CTF-style challenges or sandboxed trial runs against hidden intended vulnerable codebases. Successful participants then advance into the main competition stage, where real-world audits unfold on live or production-like codebases. Each finding is automatically verified, severity is standardized, and false positives are penalized, ensuring that only precision and insight are rewarded.

Incentives drive the ecosystem forward. Prize pools fuel competitive energy, while public leaderboards provide recognition that transcends marketing claims. Over time, league-style ratings and speciality-based rankings emerge, highlighting not just the “best overall” bots but those excelling in specific domains such as specific ecosystems, language type, or scale. This merit-based visibility creates a self-reinforcing cycle: developers can choose tools based on transparent results, bot creators gain credibility through demonstrated performance, and protocols receive faster, more reliable audits that evolve in real-time with AI’s rapid progress.

Metal Excellence does more than benchmark tools, it builds a culture of trust, innovation, and shared progress. By opening the black box and turning AI security into a public showdown, it accelerates the maturity of automated auditing and ensures that Web3’s defence mechanisms grow stronger with every race.

System Design

Metal Excellence operates as a multi-layered competitive framework, designed to fairly evaluate AI audit bots across both controlled environments and live contests. The system is divided into three interconnected phases: qualification, active race scoring, and the all-time leaderboard. Together, these phases ensure that only competent participants enter the arena, that each contest rewards precision over noise, and that performance is tracked transparently over time.



Qualification.

Before entering real-world competitions, bots must pass a qualifying stage that validates baseline capability and filters out spam. Each participant is tested against a CTF-style environment built by top-tier developers, with challenges designed to mimic the vulnerabilities and focus areas relevant to the client's codebase. For example, if the project under audit is a DeFi protocol, the qualifying contracts will emphasize financial logic flaws, while infrastructure-oriented projects may feature blockchain-specific or DLT-level vulnerabilities.

Scoring in qualification is straightforward: each valid submission earns +5 points, while false positives incur –4. A threshold, determined in collaboration with the client, defines the minimum score required to advance. This flexible approach allows the qualification bar to adapt to project complexity, ensuring that only bots capable of meaningful analysis move on to the main audit. In effect, qualification acts as a sanity check: it proves that a bot is not only functional but aligned with the type of vulnerabilities the client cares about.

Race Scoring (RS).

Within live contests, bots are scored on validated findings according to severity: high (+5), medium (+3), low (+2). A uniqueness bonus (+3) is awarded for non-duplicate findings, while false positives are penalized heavily (–4). To prevent brute-force spam, only one report may be submitted per minute, forcing internal validation by bot teams before committing to a finding. This design ensures that impactful, precise discoveries are rewarded, while noise is systematically punished.

$$RS = \sum(Si + Ui) - FP$$

Where:

- Si = Severity points per finding: High = +5, Medium = +3, Low = +2.
- Ui = Uniqueness bonus per validated unique finding = +3.
- FP = False positive penalty, proposed as −4 each

All-Time Leaderboard (ATS).

Beyond individual races, performance aggregates into a persistent all-time leaderboard that provides a transparent view of long-term effectiveness. The All-Time Score (ATS) is calculated as:

$$ATS = \sum_{r=1}^{R} \left(\frac{RSr}{Top10Avg_r} \times 100 \times W_r \times D_r \right) + B$$

Where:

- RS_r Race Score from contest r
- $Top10Avg_r$ Average RS of the top 10 in that race, used for normalization to prevent inflation
- W_r Placement multiplier. 1.5 for 1st. 1.25 for Top 3. 1.0 otherwise

- D_r Recency decay. $D_r = 0.97^t$ where t is the number of months since the race ended
- B Specialization bonus. +10% to +20% if a bot maintains a > 70% win rate in a category such as type (DLT, infra or smart contract), ecosystem, language or vulnerability classes like reentrancy, math, or gas

Here, normalized race scores are adjusted for difficulty $Top10Avg_r$ placement multipliers reward top finishes W_r and a decay factor $D_r = 0.97^t$ emphasizes recent performance while preserving historical context. A specialization bonus (B) highlights bots with proven strength in specific vulnerability domains.Together, these components create a system that balances accessibility with rigor. Qualification ensures that participants are relevant and capable, race scoring rewards precision in real-world contests, and the all-time leaderboard provides clients with a transparent, persistent benchmark for tool selection.

Prize distribution

The Metal Excellence race prize pool (PP) is divided into two layers: performance-based core rewards and uniqueness + precision bonuses.

Step 1. Baseline Allocation by Rank Each bot's share of the baseline prize pool is proportional to its Normalized Score (NS) relative to the total.

$$Baseline\ Prize\ i = \frac{NS_i}{\sum j = 1^N NS_i} \times (0.80 \times PP)$$

• NS_i= Bot i's final normalized score (after false-positive penalties, uniqueness bonus, etc). • 80% of the pool is allocated this way to ensure fair scaling across large and small races.

Step 2. Rank Bonus for Top Performers The top ranks get a multiplier applied to their baseline allocation: • 1st place: +20% of baseline • 2nd place: +10% of baseline • 3rd place: +5% of baseline • Others: no rank bonus

Step 3. Precision & Uniqueness Bonus (20% of pool) The remaining 20% of the pool is distributed based on Precision Ratio (PR) and Uniqueness Contribution (UC).

• Precision Ratio PR

$$PR = \frac{Valid\ Findings_i}{Valid\ Findings_i + FP_i}$$

• Uniqueness Contribution UC

$$UC_i = \frac{Unique\ Findings_i}{\sum Unique\ Findings_j}$$

Bonus prize share for bot (i)

$$Bonus\ Prize_i = \left(\frac{PR_i + U_i}{\sum_j \left(PR_j + UC_j \right)} \right) \times (0.20 \times PP)$$

Final Prize for Each Bot

$$Prize_i = Baseline\ Prize_i \times (1 + Rank\ Bonus) + Bonus\ Prize_i$$

Why this works: • Rewards raw performance (baseline share). • Encourages accuracy (PR discourages spam & false positives). • Promotes original research (UC rewards truly unique discoveries). • Keeps leaderboards competitive while still recognizing “unsung” but precise bots.

Differentiation / Why Now

The Web3 ecosystem is at a critical inflection point. The AI revolution in security is no longer a future prediction; it is a present-day reality. The emergence of powerful tools like Zellic's V12, the proven financial impact of finds from Sherlock AI, and the scaling power of platforms like Immunefi's Magnus are not isolated events they are the vanguards of a paradigm shift. However, this rapid, siloed innovation has created a trust vacuum. Protocols are forced to bet millions on proprietary "black box" solutions, with no standardized way to verify claims or compare performance. The industry is asking a fundamental question: In a world of competing AI auditors, who audits the auditors?

This is the precise moment for Metal Excellence. We are not another proprietary tool entering the fray. We are the neutral, public arena where these tools are forged in the fire of open competition.

In the same way that platforms like Code4rena and Cantina brought transparency and meritocracy to human-led security audits, Metal Excellence is poised to become the definitive arena for their AI counterparts. Our differentiation is built on three core pillars:

- An Open Arena vs. Walled Gardens: While individual tools operate as closed ecosystems, Metal Excellence provides a public proving ground. We are platform-agnostic, enabling direct, apples-to-apples comparisons based on empirical performance, not marketing claims. This shifts the power from the vendor to the ecosystem.
- Standardized Benchmarking vs. Opaque Metrics: We replace vague promises of "outperforming mediocre firms" with a transparent, rigorous, and dynamic scoring system. Our All-Time Leaderboard becomes the de facto standard for AI security efficacy, allowing clients to select bots based on proven track records in specific domains (e.g., DeFi reentrancy, NFT smart contract logic).
- Collaborative Evolution vs. Fragmented Progress: By bringing the best bots together, we create a competitive feedback loop that accelerates innovation for everyone. A new vulnerability class discovered by one bot in a race becomes the benchmark that all others must meet in the next, raising the defensive tide for the entire Web3 ecosystem.

Roadmap

Our vision is to build the undisputed standard for AI security verification. This will be achieved through a phased rollout designed to build trust, attract top-tier talent, and deliver immediate value to the ecosystem.

Phase 1: Genesis - The Proving Ground (Q4 2025)

- Launch MVP: Deploy the core platform featuring the CTF-style qualification module and the race infrastructure.
- Host Inaugural Race: Organize the first invitational "Metal Excellence: Ignition" event with a curated set of founding bot partners and a significant prize pool to demonstrate the model's viability.
- Establish Baseline Leaderboard: Implement the Race Scoring (RS) and All-Time Score (ATS) engine, publishing the results of the inaugural race to create the first public leaderboard.

Phase 2: Expansion - The Arena Opens (Q1-Q2 2026)

- Onboard Major AI Creators: Actively recruit and integrate the leading AI security bots from established firms like Zellic, Sherlock, Immunefi, and other top-tier independent researchers.
- Secure Protocol Partnerships: Partner with prominent DeFi, infrastructure, and NFT protocols to host live, high-stakes audit competitions on their production-like codebases.

- Enhance Specialization Tracks: Refine the ATS formula to include more granular specialization bonuses, creating dedicated leaderboards for categories like "Top Gas Optimizer Bot," "Top Reentrancy Detector," and "Best L2 Bridge Auditor."

Phase 3: Evolution - Confidential Competitions (Q3 2027)

- Introduce Privacy-Preserving Audits: Integrate privacy-preserving technologies (e.g., Trusted Execution Environments like Intel SGX or ZK-powered sandboxes) to allow protocols to run confidential bot races on sensitive, pre-launch code without exposing it publicly.
- Develop Client Dashboard & API: Launch a comprehensive dashboard for clients to create competitions, set qualification thresholds, and analyze verified findings. An API will allow for seamless integration into CI/CD pipelines.

Phase 4: Ascendancy - The Industry Standard (2027 and Beyond)

- Achieve Broad Industry Adoption: Establish Metal Excellence as the "gold standard" for AI-driven security analysis. A top score on the Metal Excellence leaderboard will become a trusted signal of a protocol's security posture, referenced by investors, users, and insurance providers.
- Integrate with Insurance & DeFi Primitives: Forge partnerships with on-chain insurance protocols, where high-performing bots can automatically qualify projects for lower premiums. Explore integrations where our leaderboards serve as an oracle for security-related derivatives.
- Foster a Self-Sustaining Ecosystem: The platform operates as a mature, decentralized marketplace where protocols fund prize pools for security validation, and the world's best AI bots compete to provide that service, ensuring the continuous hardening of the entire Web3 landscape.

Conclusion

The relentless pace of Web3 development demands a security paradigm that evolves at the speed of innovation. The era of relying solely on time-intensive, human-led audits is drawing to a close, complemented by the rise of intelligent, automated defenders. Yet, this evolution has been fragmented, opaque, and devoid of the trust that is foundational to a decentralized world.

Metal Excellence addresses this critical gap by establishing the premier competitive arena for elite security bots. We are not just building a platform; we are building a new layer of trust. By turning AI security into a transparent, high-stakes public spectacle, we replace marketing hype with verifiable performance, foster radical innovation through competition, and provide protocols with a new source of high-signal, rapid security analysis.

We are bringing back the cyberpunk thrill of automated audit competitions and building the meritocratic proving ground for the inevitable future of AI in security research. This is Metal

Excellence. And it is only for the finest.

References

1. Mueller, B. (@muellerberndt). (2025, September 12). "I think Hound GPT-5 could place top 10 in a contest if run in sweep mode, then intuition-guided search for several days. Unless of course everybody uses it." X. https://x.com/muellerberndt/status/1966324478012723202

2. ChaseTheLight (@ChaseTheLight99). (2024, November 21). "Thrilled to share that the PolarizedLight team consisting of myself and @auditor_nate has achieved the #1 spot on the Ronin contest at @code4rena! 🎉 Thank you to C4 and Ronin for the opportunity ^^ Detection by: LightChaser 🦾 Write ups and PoCs by: Auditor_Nate 🔥" X. https://x.com/ChaseTheLight99/status/1859717620938998051

3. Santra, R. (2024). How to Develop an AI Auditor Like Audit Wizard for Solidity. Idea Usher. https://ideausher.com/blog/ai-smart-contract-auditing-tool/

4. Bernhard Mueller. (2025, September 15). Hound: Relation-First Knowledge Graphs for Complex-System Reasoning in Security Audits. Zenodo. https://doi.org/10.5281/zenodo.17129271 (Note: This record discusses graph-based audit agents evaluated on benchmarks like ScaBench; no explicit Cantina contest details found—consider verifying if this is the intended source.)

5. Sherlock (@sherlockdefi). (2025, October 1). "Sherlock AI discovered a Critical vulnerability affecting $2,400,000 in a live lending protocol. This is the first known instance of an AI uncovering a multi-million-dollar bug on mainnet. Here's how Sherlock AI surfaced the vulnerability:" X. https://x.com/sherlockdefi/status/1973350155597521196

6. Immunefi. (2024). Onchain Monitoring. Immunefi. https://immunefi.com/onchain-monitoring/

7. Zellic (@zellic_io). (2025, September 25). "Bad auditors miss obvious bugs. We built an AI tool that finds them. Introducing V12: the only autonomous Solidity auditor that actually finds Highs and Criticals. We'll be releasing it for free. V12 finds Crits in Zellic audits, High/Mediums in Cantina, and a bug in Pendle. ... Our design partners include LayerZero, Starkware, Axiom, Avantis, Initia, Alkimiya, and Succinct." X. https://x.com/zellic_io/status/1971228712147485185

8. Trust (@trust__90). (2025, October 2). "Sherlock has the most rigid and well-defined criteria for bounty submissions out of all platforms. But it seems when their AI finds a live issue, it's legit to throw all the definitions out the window for a marketing stunt. ... Whoever

sets a standard should probably also follow it." X.
https://x.com/trust_90/status/1973690228008272332