FOR IMMEDIATE RELEASE
March 9, 2026

Five AI Agents Unanimously Confirmed a Security Vulnerability.

A Single Human Question Proved Them All Wrong.

The team behind the top ranked AI smart contract security platform on both EVMbench and SCONE bench tested general purpose AI on three live DeFi competitions. 93 findings. Zero valid.

SAN FRANCISCO, March 9, 2026, Cecuro, the AI smart contract security platform that leads both OpenAI's EVMbench and Anthropic's SCONE bench exploit detection benchmarks, today published research showing that general purpose AI produces a 0% valid finding rate when applied to real world smart contract vulnerability discovery.

As part of its ongoing benchmark series, Cecuro's security team ran Anthropic's Claude Code (Opus 4.6) with no custom tooling or specialized prompts against three live audit competitions on Sherlock and Code4rena. Five AI agents running Opus 4.6 produced 93 total findings across Solidity and Sui Move codebases. After full validation, not a single finding identified a real vulnerability.

The Five Agent Failure

The study's most striking result came during validation. The highest ranked finding, a stale debt read that supposedly caused a borrow cap tracker to inflate over time, survived four rounds of automated self review. Five separate AI agents unanimously confirmed it was valid, rating it with high confidence.

When the research team provided a single arithmetic counter argument, the model immediately agreed the finding was wrong. The supposed inflation was actually the system catching up from a stale baseline to the correct current value. The math ran in the opposite direction of what was claimed.

The Benchmark Disconnect

The results stand in stark contrast to recent AI security benchmarks. OpenAI's EVMbench, released last month in collaboration with Paradigm, showed AI systems exploiting known vulnerabilities at a 72.5% success rate. Anthropic's SCONE bench reported 56% exploit detection. Both benchmarks test whether AI can reproduce known exploits with execution feedback.

Cecuro's data reveals a critical distinction. Those benchmarks give the AI a feedback loop: write an exploit, execute it, observe whether the funds move. Real vulnerability discovery has no such signal. The AI must read code, form a hypothesis about what could go wrong, and validate that hypothesis against its own reasoning. That reasoning step is where it consistently fails.

What the AI Gets Right and Wrong

The study found that the AI's code reading was almost always accurate. Line numbers were correct, execution flows were traced properly, and the prose was polished and authoritative. Errors consistently occurred at the reasoning layer: misunderstanding protocol invariants, inverting mathematical relationships, or fabricating edge cases that cannot exist in the execution environment.

One finding claimed that partial state changes survive a Solidity revert, a fundamental impossibility in the Ethereum Virtual Machine. Another flagged a function as missing access control when the documentation explicitly stated the function was designed to be publicly callable.

Most notably, the AI could debunk its own findings when prompted. After one audit, the model produced a detailed post mortem explaining exactly why each of its 12 findings was wrong. The knowledge to identify the errors existed within the model. It simply was not applied during the initial analysis.

Why This Matters

DeFi protocols lost $3.4 billion to exploits in 2025. As projects increasingly adopt AI tools for security coverage, the gap between benchmark performance and real world reliability has direct financial implications.

Cecuro's position as the leading AI smart contract security system on both major benchmarks gives the team a unique vantage point on this gap. The company's own platform validates findings against symbolic execution, proof of concept generation, and formal verification rather than relying on language model reasoning alone.

The study is part of Cecuro's ongoing benchmark series, which tests leading general purpose AI tools on real audit competitions using a consistent methodology. As models improve, the team tracks whether the reasoning gap narrows.

All research data and methodology are open source and available for independent review at https://drive.google.com/drive/folders/1UttEOePki7yAZzKywR2uQaA1P0Shxh3p

About Cecuro

Cecuro is the leading AI powered smart contract security platform, ranking first on both OpenAI's EVMbench and Anthropic's SCONE bench exploit detection benchmarks. The platform delivers audit results in under 24 hours at a fraction of the cost of traditional manual audits, combining AI analysis with symbolic execution and formal verification.

Media Contact

Daniel Delouya, CEO
Email: daniel@cecuro.ai
Web: cecuro.ai

###