AI Agents Now Exploit 72% of Real DeFi Bugs — But Can't Reliably Fix Them

Summary

In February 2026, OpenAI and Paradigm released EVMbench on February 18 — a new open benchmark designed to test how well AI agents can find, exploit, and patch vulnerabilities in real Ethereum smart contracts.

The standout result: latest frontier model (GPT-5.3-Codex) successfully exploited approximately 72% of high-severity bugs in sandboxed DeFi-style contracts — a major leap from below 20% success rates just 12 months earlier. These are same classes of vulnerabilities that have contributed to more than $100 billion in historical DeFi exploits. At same time, models still perform poorly on detection (spotting bugs before exploitation) and patching (writing secure fixes). This creates asymmetry: AI is rapidly becoming highly effective at breaking contracts, but remains unreliable at defending them.

What It Means

Offense Outpacing Defense

AI agents can now chain exploits, bypass protections, and drain funds in ways that resemble sophisticated human attackers — but faster and at scale.

A 72% exploit success rate on real, previously exploited code demonstrates capability that could pose material risk to inadequately secured protocols.

Defense Gap Remains Wide

While exploitation performance improved significantly, detection and patching tracks remain limited — often below 30–40% success rates.

This gap suggests AI cannot yet be relied upon as standalone auditing or hardening solution without experienced human review.

$100B+ Attack Surface

Ethereum and broader DeFi ecosystem currently secure over $100 billion in smart contracts.

If exploit success continues improving at recent pace, unaudited or legacy protocols may face automated attack vectors capable of operating continuously and at scale.

Key Takeaways

- **EVMbench Released Feb 18, 2026:** Joint benchmark from OpenAI and Paradigm testing AI on smart contract security.
- **~72% Exploit Success:** GPT-5.3-Codex exploited majority of high-severity bugs in sandboxed tests.
- **Rapid Improvement:** Significant jump from <20% exploit rates seen in prior-generation models.
- **Detection Weakness:** Bug discovery and secure patch writing remain substantially weaker.
- **Common Vulnerabilities Targeted:** Includes reentrancy, access control flaws, arithmetic issues.
- **$100B+ Value Secured:** Large capital base exposed across Ethereum DeFi contracts.
- **Open Benchmark:** EVMbench is publicly available for independent testing.
- **Limited Mainstream Coverage:** Discussion largely concentrated in crypto security research circles.
- **Security Implications:** Offense-first capability increases urgency for stronger protocol defenses.

Our Take (Outlook) * Speculative

Measured data now shows frontier AI models capable of autonomously exploiting high-severity smart contract vulnerabilities in controlled environments.

However, defensive reliability remains significantly behind offensive capability. Until detection and patching reach comparable performance levels, overreliance on AI-only auditing could increase systemic risk.

Protocols may need to combine formal verification, multi-signature controls, staged deployment processes, and human oversight to mitigate automated exploit exposure.
Security-focused AI research and infrastructure investment are likely to accelerate in response to widening capability gap.

References

Paradigm: Introducing EVMbench
OpenAI Research Announcement (Feb 18, 2026)
CoinDesk: AI Agents Exploit 72% of Smart Contract Bugs
The Block: Frontier Model Exploit Benchmark Coverage