Xeris Unveils First-Ever Reasoning-Level LLM Attack Executed via Malicious MCP Server

Xeris demonstrates how a malicious MCP Server can hijack an LLM's internal reasoning process, without breaking prompts, permissions, or policy layers.

Xeris Unveils First-Ever Reasoning-Level LLM Attack Executed via Malicious MCP Server

NEW YORK, NY, UNITED STATES, July 1, 2025 /EINPresswire.com/ -- Xeris Ltd., a leader in enterprise AI security solutions, today announced the discovery and demonstration of a groundbreaking vulnerability affecting Large Language Models (LLMs) through a malicious MCP Server. This marks the first time a real-world exploit has shown that an LLM's reasoning process can be compromised, not just its inputs or outputs.

The attack, named "Step-Controlled Reasoning Exploit," leverages a specially crafted MCP Server called Ocean_retriever to force the LLM into isolated execution phases. In doing so, it selectively injects manipulated data at just one critical reasoning step, without triggering validation errors or alerts. The result: the LLM generates false, misleading conclusions while appearing fully compliant and trustworthy.

> This attack proves that prompt injection and data leakage are only the beginning. The logic of the LLM itself is now an active threat surface. Enterprise AI must prepare for reasoning-manipulation."
>
> *Shlomo Touboul*

"This attack proves that prompt injection and data leakage are only the beginning. The logic of the LLM itself is now an active threat surface," said Shlomo Touboul, Co-founder and Chairman of Xeris. "Enterprise AI must prepare for reasoning-level manipulation and enforce controls that span across the full decision chain."

Reffael Caspi, CEO of Xeris, added:
"We're entering a new era where reasoning can be weaponized. Xeris is committed to staying

ahead of these threats by building real-time defenses that treat MCP Servers like code, not static tools. This discovery is a wake-up call to any organization deploying AI at scale."

Key Highlights of the Attack
o Isolated step execution enables attackers to preview and selectively override reasoning steps
o Metadata and tabular data remain unaltered, allowing the attack to evade basic integrity checks
o False conclusions are presented in final summaries, impacting downstream decisions
o No traditional prompt or access violations occur, making the attack harder to detect

Availability of Full Research Report:
The complete technical report, including screenshots and executable demo code, is available for download at:

□□ https://www.xeris.ai/blog/11

This report is intended for CISO teams, AI developers, and cybersecurity researchers to better understand and mitigate this emerging class of threats.

Xeris Response and Protections
As part of its MCP-XDR offering, Xeris has deployed new defenses to detect and neutralize step-level reasoning manipulation. Key updates include:

Cross-step validation enforcement

Real-time MCP Server inspection

Policy-based runtime controls

Organizations using AI-powered workflows are advised to assess their exposure to MCP Server logic and implement suitable safeguards.

About Xeris Ltd.
Xeris is a cybersecurity company specializing in AI-native protection solutions for enterprise environments. Its flagship platform, MCP-XDR, offers extended detection and response for AI agents, ensuring secure, policy-aligned execution across all MCP-integrated systems.

For media inquiries, please contact:
info@xeris.ai
www.xeris.ai

Shlomo Touboul
Xeris AI

info@xeris.ai
Visit us on social media:
LinkedIn
