AI Is Hacking You: The Rise of Agentic AI in Offensive Security

Dhillon @l33tdawg Kannabhiran

(VP Global Strategy & Growth @ Verichains / Founder @ HITB)

Who Am I?

Founder of Hack In The Box (**the security conference**)

Currently also VP of Global Strategy & Growth @ Verichains

Music producer and DJ during my spare time

Vibe coder of:

- SilentGem (a transparent Telegram translator using AI)
- PhantomRecon (AI agent-driven red team automation tool)

Follow me on:

Twitter: @l33tdawg

Instagram: @dhankasounds

Github: https://github.com/l33tdawg/

DJ / Music stuff: https://linktr.ee/dhankasounds/

The AI Security Landscape is Changing Every Day

AI as Attacker

AI is no longer just a chat bot — it's becoming the attacker itself.

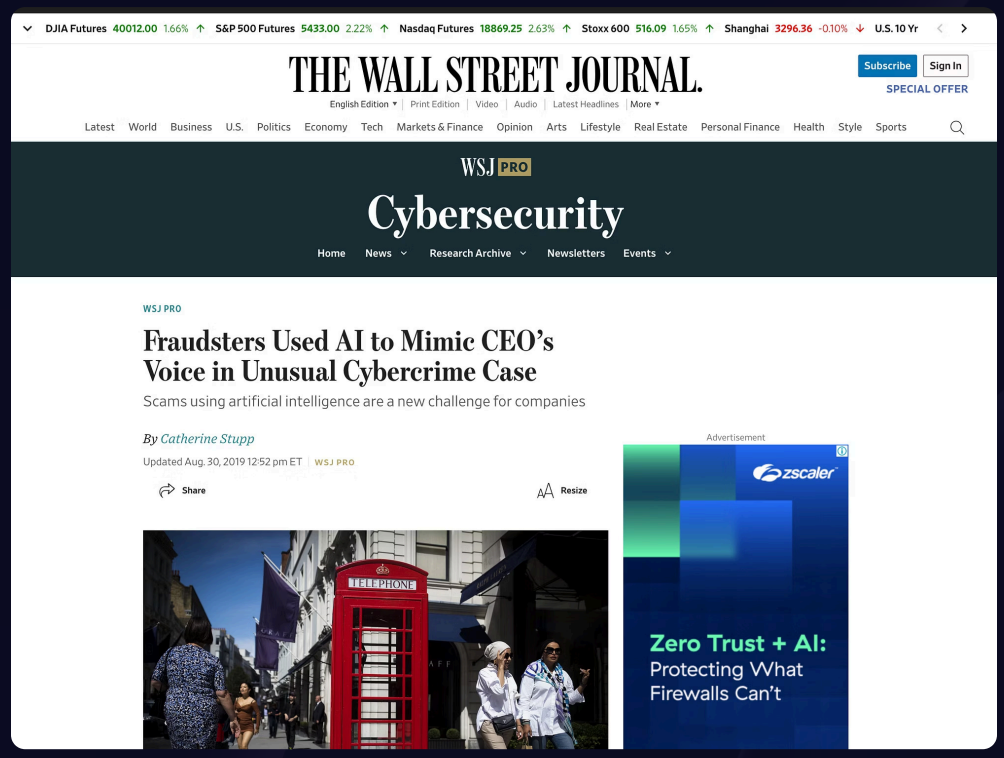
Autonomous Threats

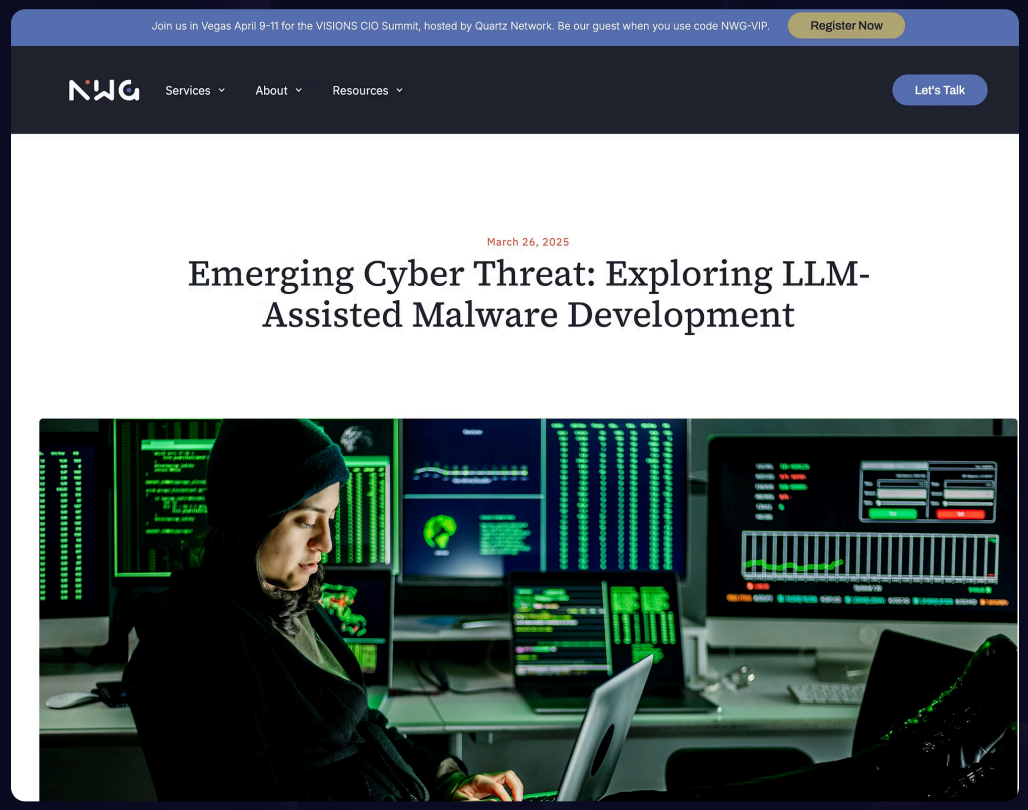
Autonomous malware, phishing bots, and LLM-assisted reconnaissance in the wild.

Evolving Attacks

Attacks are getting faster, weirder, and definitely harder to trace.

Real-World Examples





FraudGPT/WormGPT Clones

Used for spear phishing, fake identities, and malware creation.

Voice Cloning Used for Real-Time Phone Scams

Attackers cloned a company executive's voice to request urgent wire transfers from finance staff. Deepfakes + LLM scripting used to call IT helpdesks with "urgent reset requests" in cloned voices of executives.

LLMs Used for Polymorphic Malware Generation

Attackers use LLMs (like GPT-4 or Gemini Pro) to generate **code variants of known malware** with each execution.

These variants change function names, obfuscate logic, reword strings, and even alter protocol behaviors — **all without changing the payload's core behavior**.

AI as the Attacker

Human Attackers

- Logic-based approaches
- Slow, manual processes
- Limited creativity
- Needs to sleep

AI Attackers

- Introspective & creative
- RAG → CAG → Multi-agent systems (Google ADK)
- Can 'think' and plan
- Can write code - **fast**
- Adapts to defenses
- Scales infinitely

What Can You Start Hacking Right Now?

Prompt Hacking

Try jailbreaking your own GPT to understand vulnerabilities around things like chat guardrails and how they can be bypassed to **make GPTs do 'evil things'** (e.g. gain of function research that uses genetic modifications to enhance the capabilities of viruses and pathogens - COVID9000)

Vibe ~~Coding~~ Hacking

Describe your problem, your goals, and be specific on what you'd like to achieve - use voice input; **it's faster than typing**

Tools to Try

- ChatGPT-4.1, Claude 3.7, Gemini 2.5 Pro Exp
- **Google Agent Development Kit -** a flexible and modular framework for developing and deploying AI agents.
- **LangChain** – orchestration for AI agents and tools.
- **Autogen (Microsoft)** – multi-agent planning + tool execution.
- **CrewAI** – lightweight agent framework ideal for offensive workflows.
- **LLamaIndex** – great for building local knowledge base agents (like a CVE searcher bot).
- **Burp Suite + Copilot (BurpGPT plugin)** – generate & test payloads in real time.
- **pwntools + GPT** – combine for exploit generation.
- **ReconFTW + LLM assist** – automate recon with natural language layer.
- **Nuclei + LLM Templates** – build hybrid scans with generated payloads.

Real-World Vibe Hacking

1

Describe Goal

"Find a SQL injection point in this PHP snippet"

2

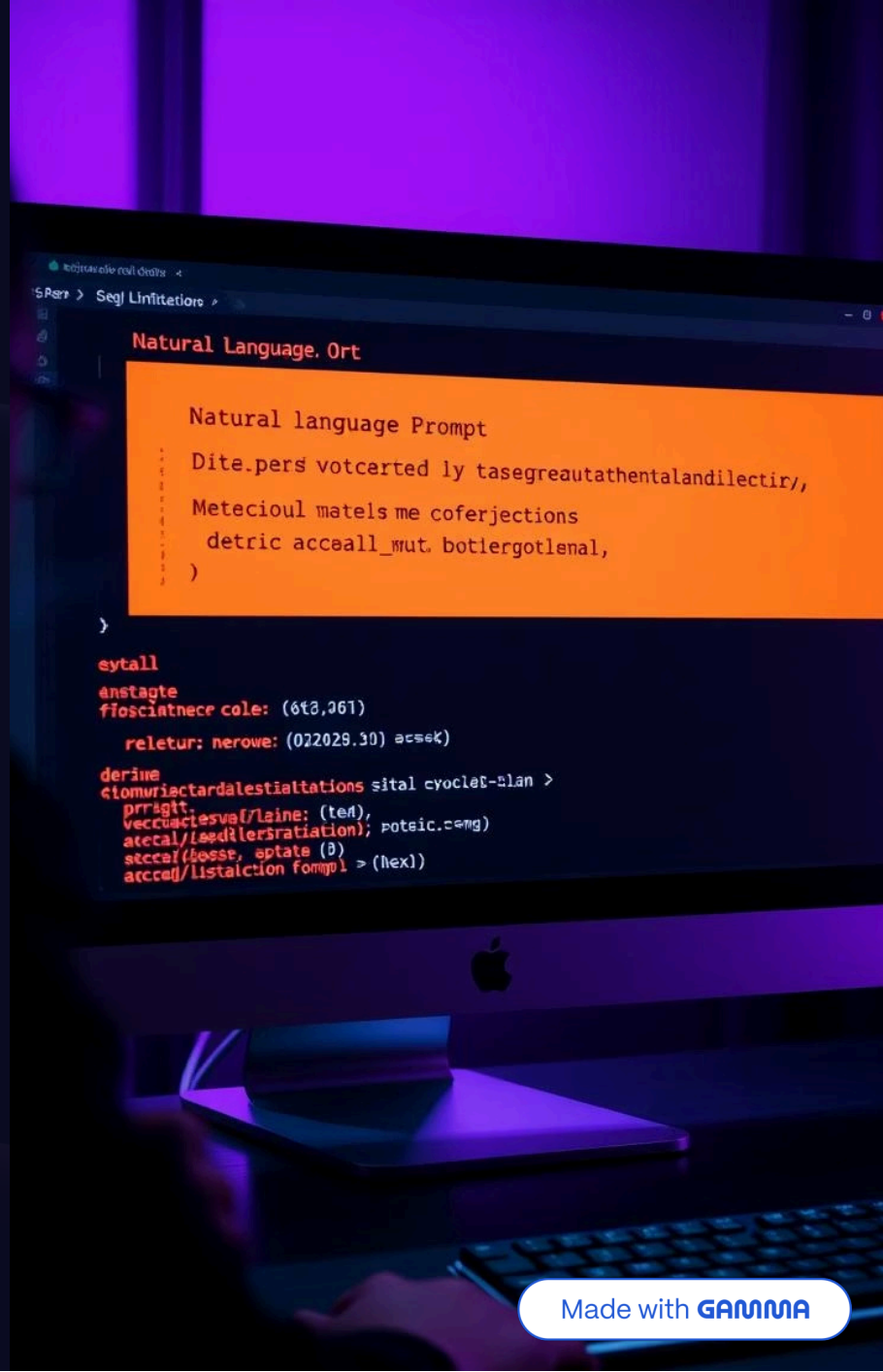
Generate Payload

"Exploit SQLi login form to bypass authentication"

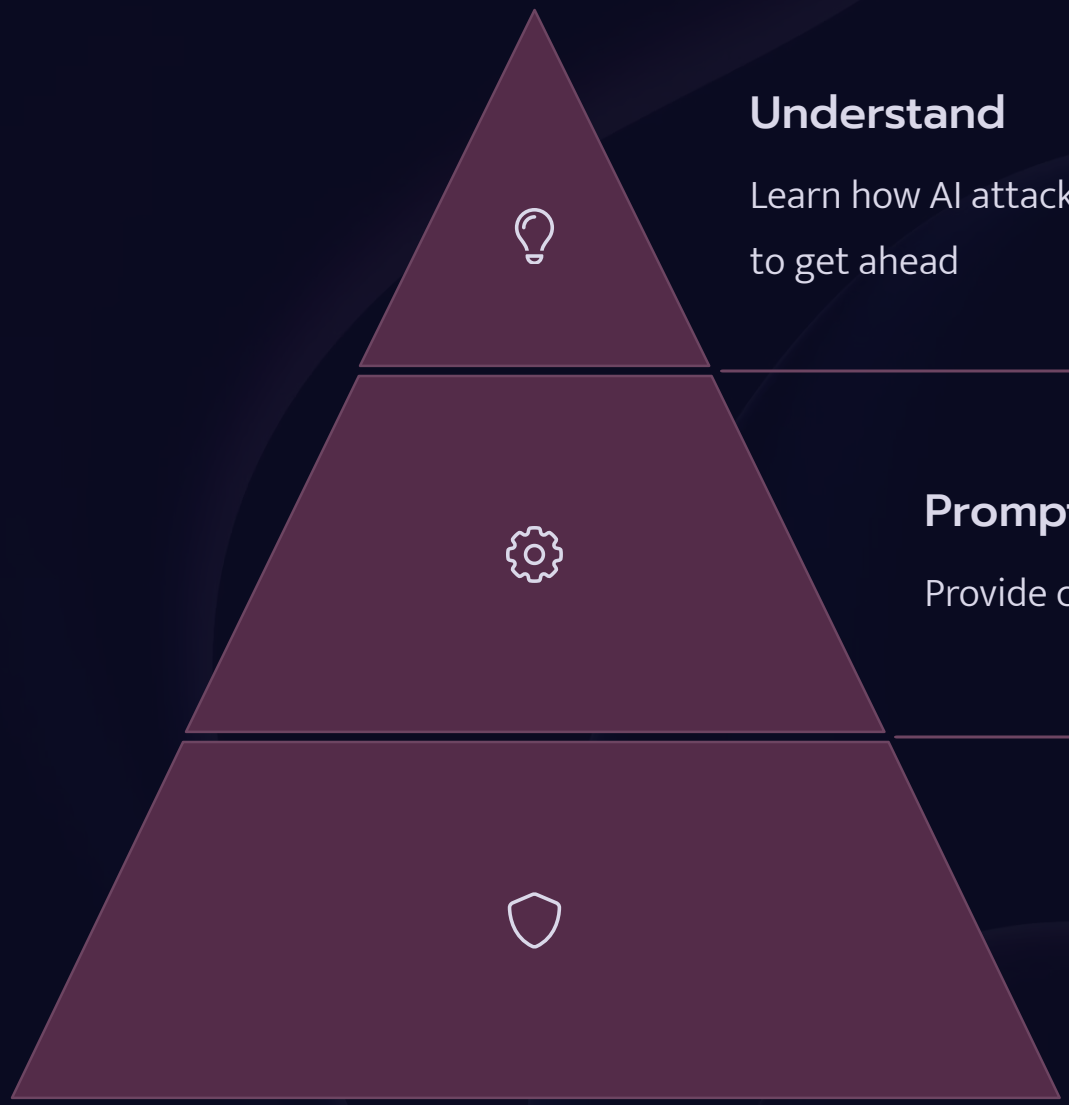
3

Execute & Reiterate

Focus on 'attacker thinking' prompts beyond 'help me break this'



The Future Belongs to Builders & Breakers



Understand

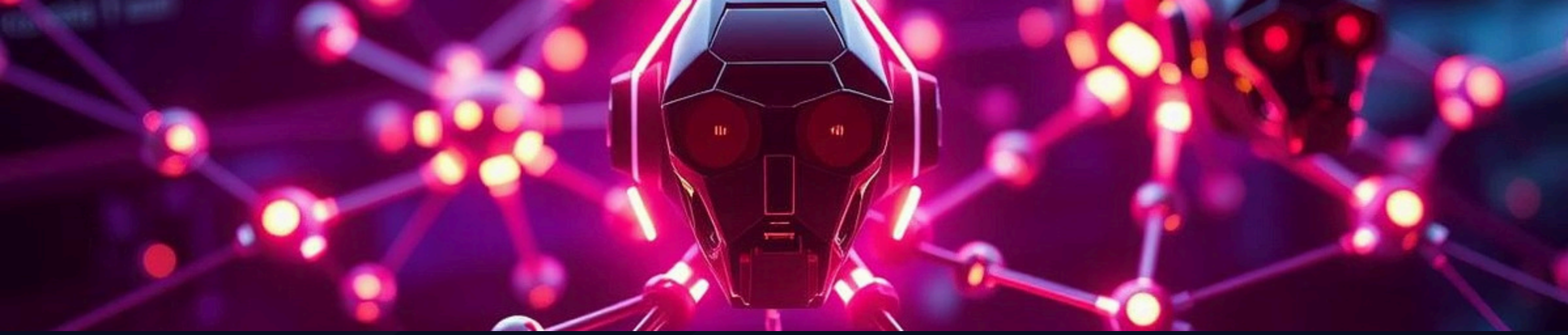
Learn how AI attacks work and how to leverage the latest tools and frameworks to get ahead

Prompt KungFu

Provide clear context and goals - make use of things like rules for IDEs

MCP and Agents for the win

Make use of Model Context Protocol to extend 'knowledge' and functionality. Build specialized agents and orchestration workflows.



How an Agentic Attack AI Works

Recon Agent

Gathers intelligence on targets

Planner Agent

Develops attack strategy

Exploit Agent

Executes vulnerabilities

Report Agent

Documents findings

"Agentic" means multiple steps with goal planning. These are no longer scripts—they're adaptive task forces.

PhantomRecon: Your Own Agentic Red Team

l33tdawg/
phantomrecon

PhantomRecon is a CLI-based, modular, agent-driven red team automation tool designed to demonstrate autonomous offensive security workflows powered by AI...

1 Contributor | 0 Issues | 0 Stars | 0 Forks

GitHub

GitHub – l33tdawg/phantomrecon: PhantomRecon is a CLI-bas...

PhantomRecon is a CLI-based, modular, agent-driven red team automation tool designed to demonstrate autonomous offensive securit...

Scan Target

Automated reconnaissance of the target system

Plan Attack

AI strategizes based on discovered vulnerabilities

Select Exploit

Choose optimal attack vector

Delegate to Agents

Specialized AI agents execute specific tasks

Generate Report

Produce human-readable security findings

PhantomRecon Architecture

Web & CLI Interface

Launch from terminal or via adk web

Reporting Agent

Documents findings and outputs report in PDF and HTML

Recon Agent

Gathers target intelligence from DNS, nmap, web

Planner Agent

Strategizes attack vectors based on findings from Recon Agent

Exploit Agent

Executes planned attacks using installed tools like sqlmap or metasploit

Modular agent design works like red team Lego blocks. Each agent is a standalone unit that can be replaced or upgraded or expanded with further capabilities

PhantomRecon - How You Can Extend It

1

Clone the repo

https://github.com/l33tdawg/phantomrecon/

2

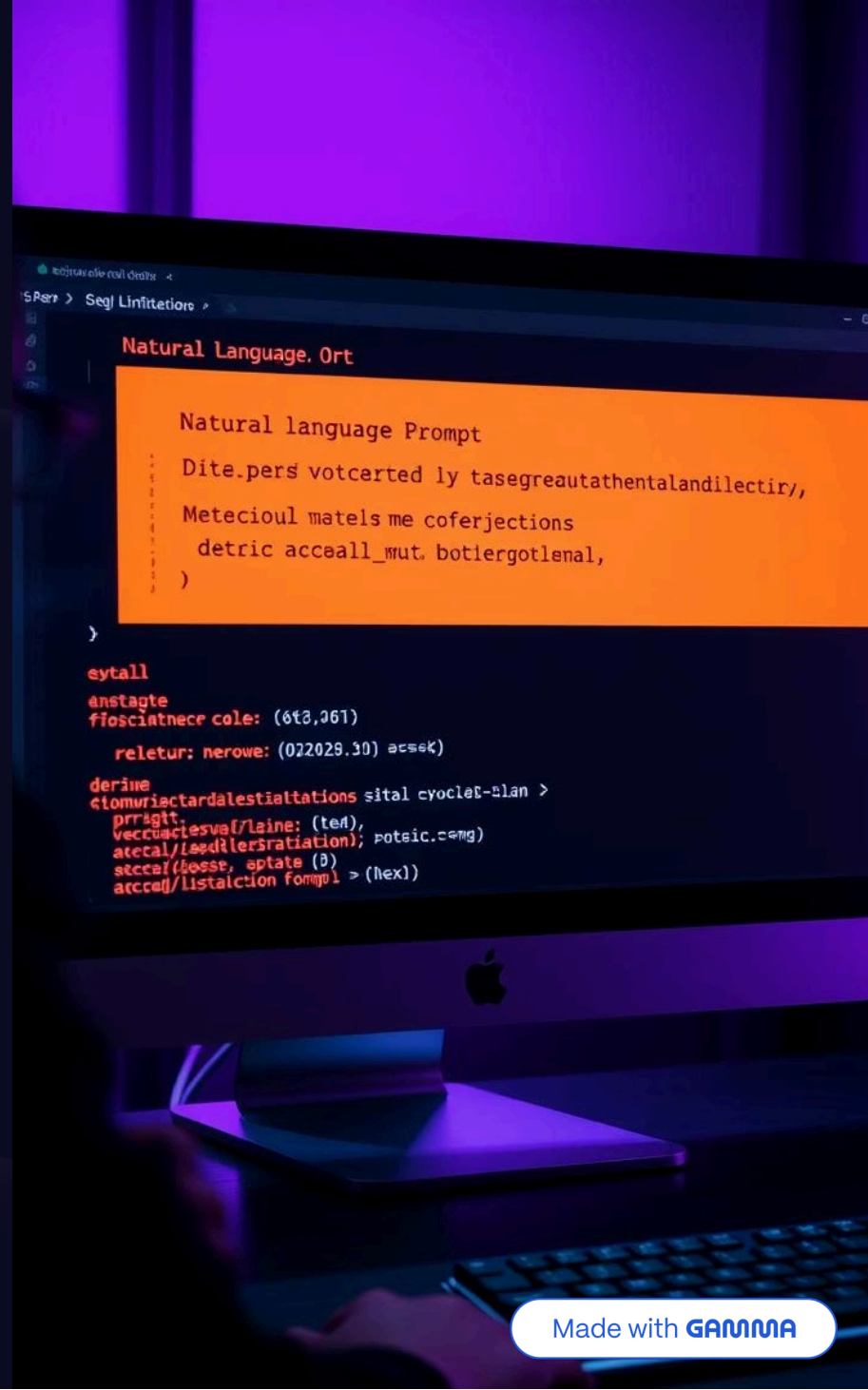
Connect it to your IDE of choice

Cursor, Cline, Windsurf etc

3

Add Your Own Agent

Prompt your way to extending the functionalities available



Thank You For Your Time - Now Go Build Something Awesome!

Email: dhillon@verichains.io

Twitter: @l33tdawg