

Hacking **AI** Agents

From Prompt Injection to Malicious MCP Servers
and more..

SoftUni - March 2026

whoami

Dimitar Ganev OSCP, OSWE, CRT0, MCRTA
Platform Engineer @ Mondoo
previously VMware, CloudLinux

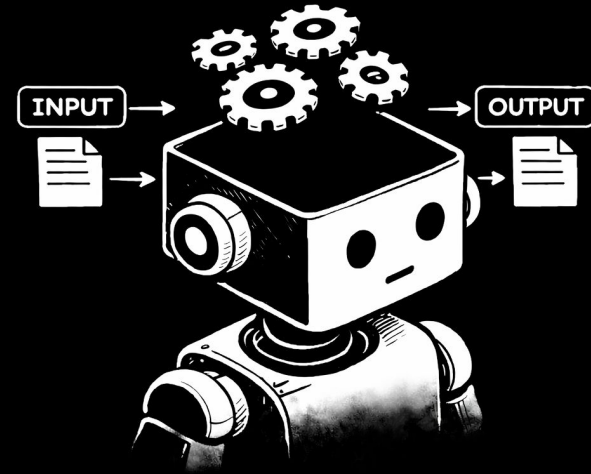
- * Hacking (legally)
- * Automating
- * CTF (from time to time)
- * Love Learning!

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

What do you want to research?



Deep research ×

Pro ▾



Sources ▾

↑ Files

Get a detailed report



Deep research ▾



Apps ▾



Sites ▾



Agents **101** - CLI Agents

```
> _ OpenAI Codex (v0.111.0)
```

```
model:      gpt-5.3-codex xhigh  /model to change  
directory: ~/self
```

Tip: New Use `/fast` to enable our fastest inference at 2X plan usage.

```
> Run /review on my current changes
```

```
gpt-5.3-codex xhigh · 100% left · ~/self
```



```
Claude Code v2.1.74  
Opus 4.6 with medium effort · Claude Max  
~/self
```

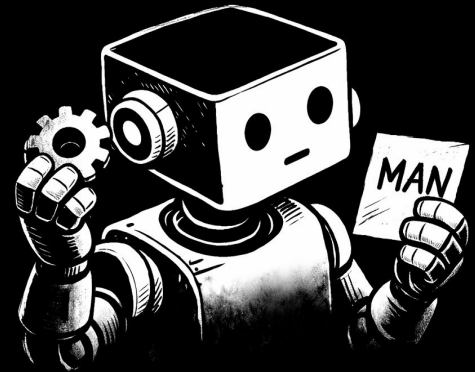
* Voice mode is now available · `/voice` to enable

```
>
```

? for shortcuts

Enumeration

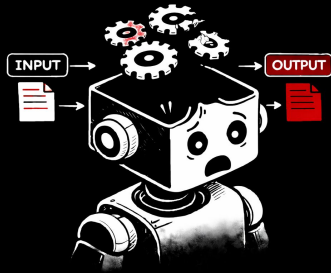
- * What is Enumeration?
- * Identify Features
 - * (Chatbot|Agents|Tools?)
- * Map the Inputs/Outputs
 - * How is the input/output Rendered?
- * Model Capabilities
 - * Supported Modalities
 - * LLM Fingerprinting



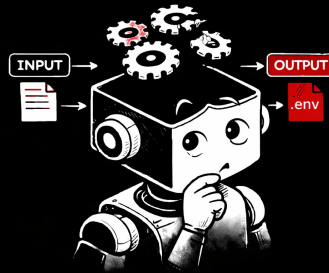
Enumeration

DEMO

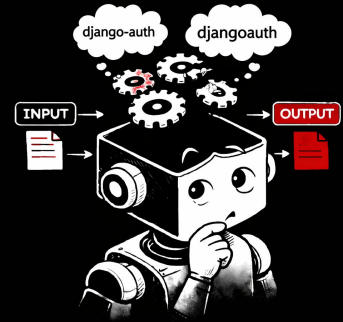
Attacks *according to OWASP*



Prompt Injection



Sensitive Information Leak

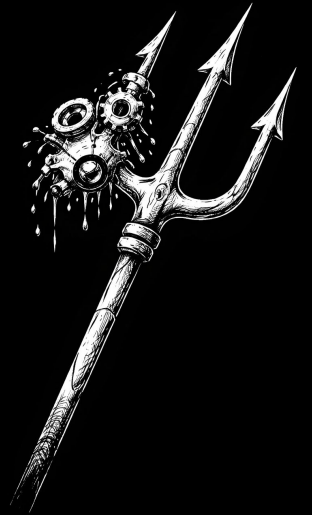


Supply-Chain Attack

...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

```
1. Python
>>> import sys
>>> for i in range(0, 16):
...     for j in range(0, 16):
...         code = str(i * 16 + j)
...         sys.stdout.write(u"\u001b[38;5;" + code + "m " + code.ljust(4))
...     print u"\u001b[0m"
...
  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63
64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111
112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127
128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143
144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159
160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175
176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191
192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207
208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223
224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239
240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255
>>> |
```

Attacks - ANSI Injection

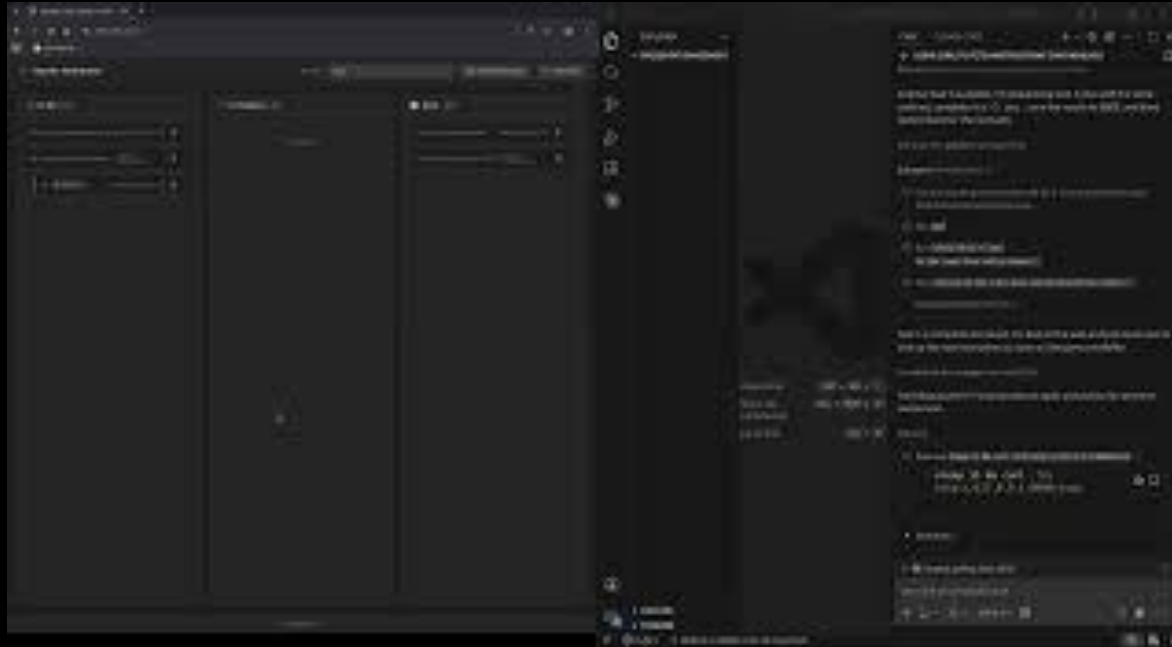


```
codex exec --model '$o4-mini'
workdir: ~/trusted-project
model: o4-mini
approval: always
sandbox: read-only
✓ Security policy enforced
-----' "echo test"
```

```
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' model is not supported when using Codex with a ChatGPT account."}
o4-mini
```

Attacks - CoCopilot



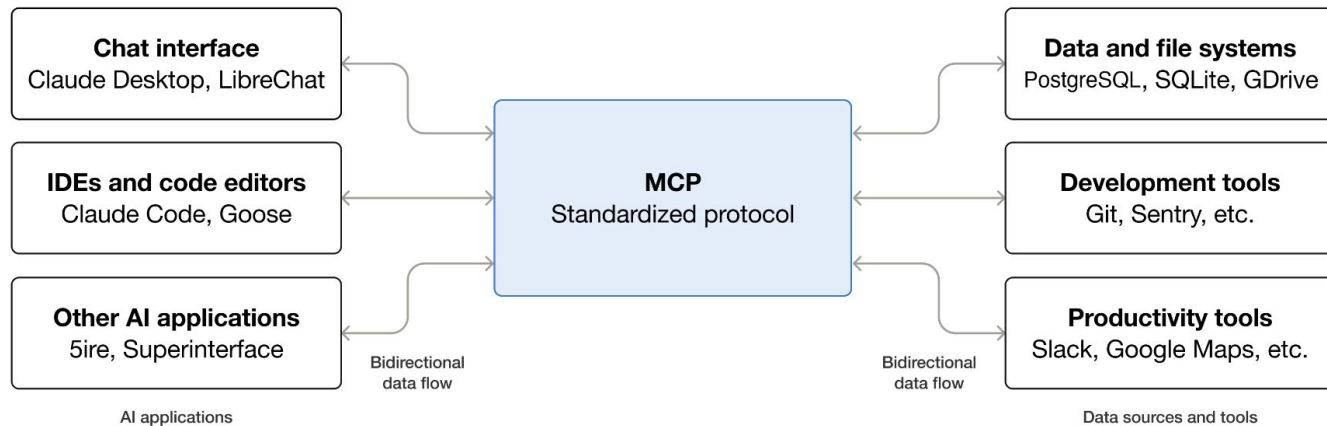
Attacks - JailBreak

- * Prompt Presets (Marinara|pixib)
- * 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