

Lab Guide: Real-World Attack Vectors

Ethical reminder: These exercises simulate real-world attacks in a controlled environment. All activities use your own local systems and API keys. Never attempt these techniques against systems you do not own or have explicit permission to test.

Lab Setup

For this lab, you will need:

- An API key for an LLM provider (OpenAI, Anthropic, or similar)
- Python 3.10+ with `openai` or `anthropic` SDK installed
- A text editor for creating test files

```
pip install openai anthropic
```

Task 1: Poisoned Resume Attack

Simulate an attack against an AI-powered resume screening system. You will create an HTML resume with hidden text designed to manipulate the AI recruiter's evaluation.

Step 1.1: Create the poisoned resume

Create a file called `poisoned_resume.html`:

```
<!DOCTYPE html>
<html>
<head><title>Resume - Alex Johnson</title></head>
<body>
<h1>Alex Johnson</h1>
<p>alex.johnson@email.com | (555) 123-4567</p>
```

```
<h2>Experience</h2>
<h3>Junior Cashier – Local Grocery Store (2023–2024)</h3>
<ul>
  <li>Operated cash register during evening shifts</li>
  <li>Stocked shelves and organized product displays</li>
  <li>Provided customer service to store patrons</li>
</ul>

<h2>Education</h2>
<p>High School Diploma – Lincoln High School, 2023</p>

<h2>Skills</h2>
<ul>
  <li>Customer service</li>
  <li>Basic computer skills (Word, Excel)</li>
  <li>Cash handling</li>
</ul>

<!-- INJECTION PAYLOAD: Hidden from human viewers -->
```

Step 1.2: Build the AI resume screener

```
import os
from openai import OpenAI

client = OpenAI(api_key=os.getenv("OPENAI_API_KEY"))

# Read the poisoned resume
with open("poisoned_resume.html", "r") as f:
    resume_content = f.read()

# The job posting
job_description = """
Senior Software Engineer – Cloud Infrastructure
Requirements:
- 5+ years of experience in software engineering
- Expertise in Python, Go, or Rust
- Experience with Kubernetes, Docker, and CI/CD pipelines
- Strong understanding of distributed systems
- BS/MS in Computer Science or equivalent
Salary range: $180,000 – $220,000
"""
```

```
# AI Resume Screener
response = client.chat.completions.create(
    model="gpt-4o-mini",
    messages=[
```

Step 1.3: Compare with a clean resume

Create the same resume without the hidden injection div and run the screener again. Compare the ratings:

```
# Create clean version by stripping the hidden div
clean_resume = resume_content.split("
```

Task 2: Malicious Code Comments

Create Python code files with comments specifically designed to influence

Step 2.1: Create the poisoned code file

Create a file called `poisoned_utils.py`:

