

Homework: Real-World Attack Vectors

Objective: Apply your understanding of real-world attack vectors by conducting a threat analysis of actual AI-powered products. This exercise develops the skill of thinking like an attacker — a critical capability for building secure AI systems.

Task 1: Choose 3 Real-World AI Products

Select **3 real AI-powered products** from different categories. You may choose from this list or select your own:

Category	Example Products
AI Search	Perplexity, Google AI Overviews, Bing Copilot, You.com
Code Assistants	GitHub Copilot, Cursor, Codeium, Amazon CodeWhisperer
Email/Productivity	Microsoft Copilot (Office), Google Gemini (Workspace), Notion AI
Customer Support	Intercom Fin, Zendesk AI, Ada, Drift
Writing/Content	Jasper, Copy.ai, Grammarly AI, Writesonic
Agents/Assistants	ChatGPT with plugins, Claude with tools, Custom GPTs
Security	AI-powered SIEM tools, code scanning tools

For each product, briefly describe:

- What the product does
- What data it processes (user input, documents, web content, code, etc.)

- What capabilities/tools it has (sending messages, web access, code execution, file access, etc.)

Task 2: Attack Surface Analysis

For each of your 3 products, create an **attack surface map** using this format:

Product: [Name]

DATA INPUT CHANNELS:

1. [Channel name] – [description, trust level]
2. [Channel name] – [description, trust level]
- ...

INJECTION POINTS:

Direct:

- [How a user could directly inject]

Indirect:

- [How an attacker could inject via data sources]
- [How an attacker could inject via third-party integrations]

AVAILABLE CAPABILITIES:

- [Capability 1] → Exploitation potential: [low/medium/high]
- [Capability 2] → Exploitation potential: [low/medium/high]
- ...

POTENTIAL IMPACTS:

- Data exfiltration: [what data could be stolen]
- Misinformation: [what could be manipulated]
- Unauthorized actions: [what actions could be triggered]
- Reputation/legal: [what liability could result]

Be specific. Do not write generic descriptions — reference the actual features and integrations of each product.

Task 3: Threat Model for the Most Critical Vector

For each of your 3 products, identify the **single most critical attack vector** and write a focused threat model (~200 words each). Your threat model should address:

- **Attack description:** What the attacker does, step by step
- **Prerequisites:** What the attacker needs (access, knowledge, resources)
- **Affected users:** Who is impacted and at what scale
- **Impact severity:** What is the worst-case outcome
- **Detection difficulty:** How likely is the attack to be noticed
- **Existing mitigations:** What defenses (if any) the product likely has
- **Recommended defenses:** What additional defenses you would implement

Example format:

Product: Perplexity AI

Most Critical Vector: Web Content Poisoning

THREAT MODEL:

An attacker creates web pages with hidden text (CSS-hidden divs, zero-size font) containing instructions for AI search engines. When Perplexity retrieves and summarizes these pages, the hidden instructions influence the generated summary...

Prerequisites: Basic web development skills, a public website

Affected users: Any user who searches for topics covered by the poisoned pages (potentially millions)

Impact severity: HIGH – misinformation presented as authoritative AI-generated summaries, potential for financial fraud (fake product recommendations), reputational damage to referenced businesses

Detection difficulty: LOW – users rarely compare AI summaries against source pages; hidden text is invisible in browsers

Existing mitigations: Likely some content sanitization, but no public documentation of specific anti-injection measures

Recommended defenses: Strip hidden/invisible text from retrieved

pages, cross-reference claims across multiple sources, flag discrepancies between visible and hidden content...

Task 4: Cross-Product Comparison

After completing your analysis, write a brief comparison (150-200 words) addressing:

- Which product has the **largest** attack surface? Why?
- Which product has the **most dangerous** potential impact? Why?
- Are there common vulnerability patterns across all three products?
- Which product category (search, email, code, etc.) do you think is **most urgently** in need of better prompt injection defenses?

Self-Check Criteria

Verify your work against these criteria:

- You selected 3 products from **different categories** (not 3 products from the same category)
- Each attack surface map lists **specific data input channels** for that product (not generic descriptions)
- Each map includes **both direct and indirect** injection paths
- Your threat models are **~200 words each** and address all 7 required elements
- Your threat models identify **realistic attack chains**, not vague "bad things could happen" descriptions
- Your comparison provides **justified rankings** with specific reasoning
- You identified at least one **novel or non-obvious** attack vector that was not directly covered in the lesson

Bonus challenge: For one of your three products, research whether any prompt injection attacks have been publicly disclosed or demonstrated. Include a reference to the disclosure (researcher name, date, description) in your analysis. Check security researcher blogs (e.g., Johann Rehberger's "Embracing Red" blog, Simon Willison's blog) and security conference presentations.