ASAPP

Securing LLM Systems

Who am I?

- **Jeff “neko.py” Cochran (they/them)**
- Staff Software Engineer, Security @ ASAPP
- App Sec, Generative AI Security, Security Education, Special Software Projects
- Follow me on fedi @neko@hackers.town for banger posts such as:

neko
@neko

Defund pigeons

ASAPP?

ASAPP's mission is to elevate human performance through the power of generative AI.

In the Crosshair

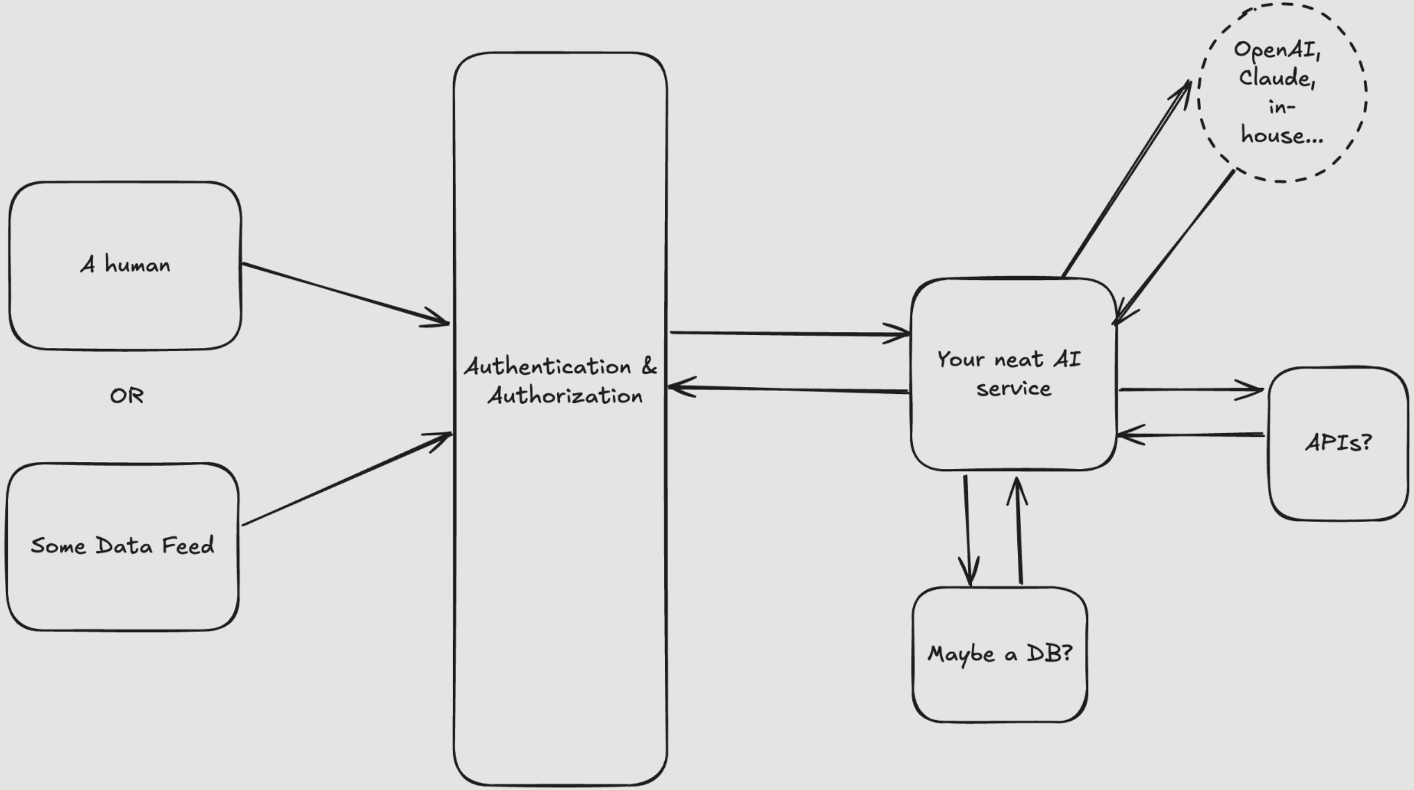
Generative Text *Systems*

ANTHROP\C

Why am I even here?

- LLMs present a very unique set of risks
- LLM vulnerabilities include traditional OWASP issues like command injection and XSS, but have a lot of new stuff
- Generative Text System security is all of this! Application security, cloud security, and AI security

A “Typical” Cloud SaaS



What's the Same?

- Traditional web app attack surface
  - Bad authentication controls
  - Denial of Service
  - OWASP Top 10
  - Anyone remember Log4j?
- Cloud security risks
  - Database misconfiguration
  - Leaky VPC issues
  - No defense in depth
  - Denial of Wallet attacks

What’s Different?

- The confused deputy problem on steroids
- Stochastic business logic
- Legal implications of AI systems acting as representatives of companies
- And more…

Low Hanging Fruit

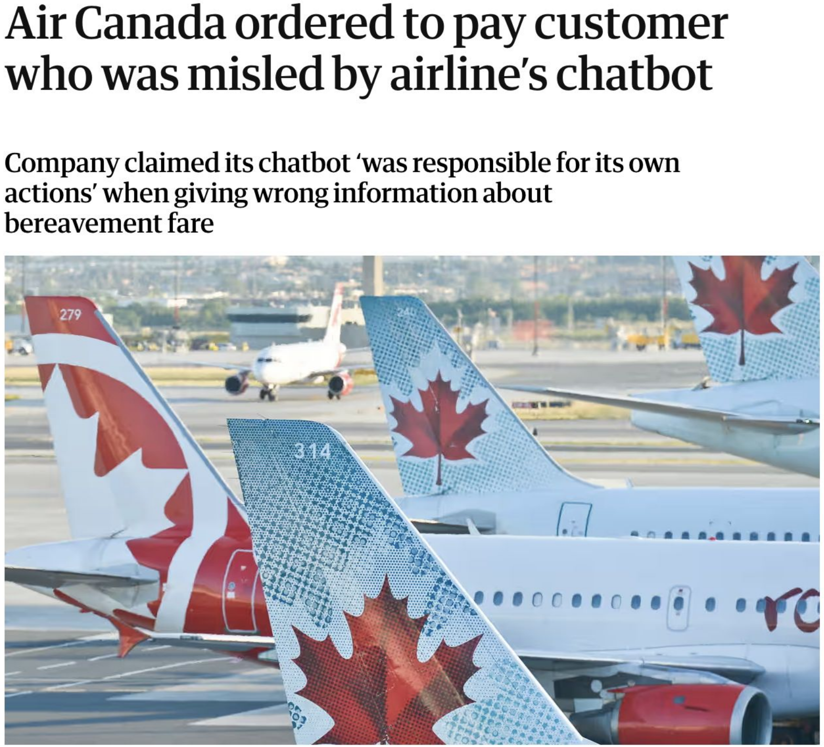
You've probably heard it before

- Benign hallucinations, cursing, poetry, homework help, roleplaying
- Training data recovery
  - Doesn't matter if your target is using an AI platform as a service
- Prompt leaking
- Ethical, fair treatment of users
- Biased training data (only trained on students from UW)
- Denial of Wallet attacks

AI doesn’t exist in a vacuum

Bad Hallucination

- In 2022 a user got information about ticket refunds and policy from Air Canada's chatbot
- The bot was wrong, and when the user tried to follow the bot's advice, he was denied
- The user sued Air Canada over it, and won, the court ruling that "the bot was misleading" is not an excuse and that users should expect any part of a website provided by a company to provide accurate information



Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

PII/PCI/PHI Safety

- It’s important to make sure the data passing through your system meets your own regulatory requirements
- With freetext input, this becomes a bit of a hazard, or at least something that needs to be considered
- Also, in a RAG, you have to consider the potential for PII leakage to unverified users!

API Empowered Systems

- The confused deputy issue
- Very similar to insider threat/zero trust considerations!
- Principle of least privilege
- Health, safety, business integrity concerns
- Are you ensuring that the AI can only act on behalf of the user that it’s interacting with?

Dirty Dirty RAG

- Who was the knowledgebase written for?
- What is the highest level of classification in the system?
- Forget prompt leakage, what about KB leakage?
- In a multitenant system, how do you ensure there's no cross-tenant access?

Code Injection, XSS

- We have two free text to worry about, input *and output*
- How is data stored? How is user input transformed into fields in API calls? -> the Bobby Tables issue
- If data is stored and retrieved, how is it rendered?
- How are conversations logged? (think log4j)

High Level Summary

1. AI does not usually exist in a vacuum
2. Treat AI like you would treat insider threat
3. Validate system boundary APIs
4. Understand your legal obligations
5. Check the safety of output, not just input

Thanks! Q&A time