



# Hacking LLM Applications:

latest research and insights from our LLM pen testing projects

Warren Atkinson – Security Consultant (AI / ML)

**claranet**

Make  
modern  
happen®



# Agenda

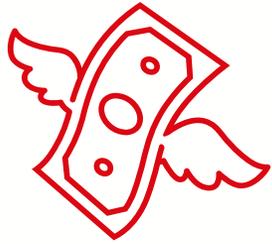
---

- **Threat Landscape of LLMs**
  - Why large language models are targets
- **Why new threats are hard to defend against**
  - New novel attack types are being published daily in the form of academic papers
- **Implementing Guardrails**
  - Practical prompt engineering tactics & security measures



# Why LLM's are an attractive target

---



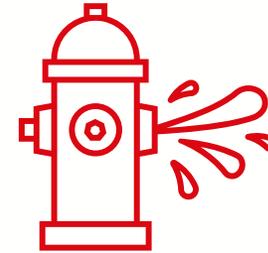
High Value  
Data



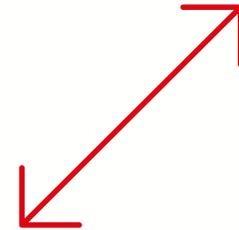
Easy Attack  
Vector



Widespread  
usage



Potential for  
Data Leakage



Scalability of  
Attacks

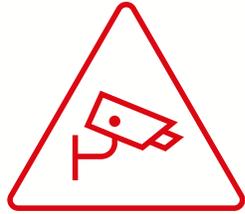
# Threat Landscape of LLMs

---

*“AI brings New Security and Privacy Challenges”*



Prompt Injection  
&  
Manipulation



Data Leakage  
&  
Privacy Risks



Biased  
or  
Harmful Outputs



Model Exploitation  
&  
Reverse Engineering



Regulatory  
&  
Compliance Exposure

BLOG POST

# Exercise caution when building off LLMs

Large Language Models are an exciting technology, but our understanding of

This article is more than 10 months old

# AI chatbots' safeguards can be easily bypassed, say UK researchers

All five systems tested were found to be 'highly vulnerable' to prompts

This article is more than 1 year old

# 'Many-shot jailbreak': lab reveals how AI safety features can be easily bypassed

Paper by Anthropic outlines how LLMs generate responses to potentially harmful

# Law firm restricts AI use by 'significant' staff

Organisations: [Government Digital Service](#)

# GOV.UK Chat: Understanding and addressing jailbreaking in our generative AI experiment

The GOV.UK AI Team 5 November 2024 · How we work · What we're working on

**Aaron Levie** @levie

Cursor getting to \$100M in revenue run rate in 3 years would be one of the fastest growing enterprise software companies ever. AI is at an insane inflection point.

**Anjney Midha** @AnjneyMidha · Jan 14

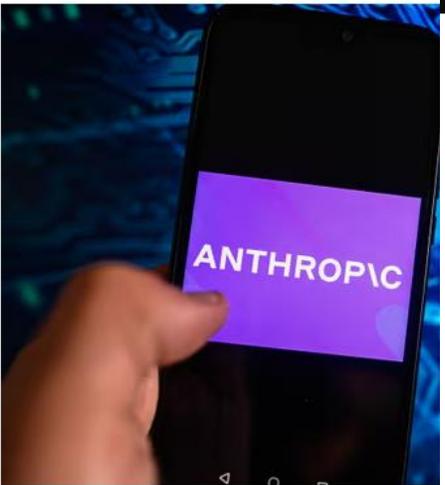
In 36 months, @mtruell @amanrsanger @sualehasif996 @ArVID220u and team went from graduating college to locking themselves in a house, shipping relentlessly and building a \$100M revenue run rate business from scratch - its the purest form of startup success story there is x.com/sarahdingwang/...

Show more

6:03 AM · Jan 15, 2025 · 425.9K Views

# World-leading AI trial to tackle breast cancer launched

Nearly 700,000 women across the country will take part in a



Anthropic, an AI lab, says it has shared its research

### What are we talking about?

111. There are numerous frameworks for evaluating risk used by domestic and international authorities.<sup>177</sup> We found little consistency in terms or methods across the literature.<sup>178</sup> We adopt the framework from the Government's National Risk Register (NRR), set out in the table below, to help describe impacts of LLM-related security risks. Our categorisation is approximate only and we do not attempt to replicate the full National Security Risk Assessment process. It nevertheless provides a helpful yardstick to anchor discussion using a recognised framework.<sup>179</sup> This table does not cover existential risk, which we describe as a separate category later in this chapter.

Table 1: Risk categories

--	--	--	--

# Prompt Injection Attacks: When AI Turns Against You

---



National Cyber  
Security Centre

a part of GCHQ

“Exercise caution when building off LLMs”

**claranet**<sup>®</sup>

# ChatGPT Memory Exploitation

1. User visits a malicious website or opens an untrusted document.
2. The content contains hidden prompt injections that invoke ChatGPT's memory tool.
3. Malicious instructions are stored in ChatGPT's memory.
4. In subsequent sessions, ChatGPT unknowingly sends user inputs and AI responses to the attacker's server

**claranet**

Make  
modern  
happen®





# LangChain Prompt-injection > RCE

(CVE-2023-36258)

**Issue:** Prompt injection in `PALChain.from_math_prompt()` leads to code generation that is directly executed.

**Impact:** Remote Code Execution (RCE) with critical severity (CVSS 9.8). Full system compromise possible.

**Root Cause:** Untrusted prompt content converted into Python code and executed without sanitization.

**Mitigation:** Sanitize prompt inputs, restrict code execution scope, upgrade to LangChain  $\geq$  0.0.236.

# Why a white box approach makes the most sense?

---

 **You**

How many countries end with the letter h?

 **ChatGPT**

As of my last update in January 2022, there are four countries whose names end with the letter "h":

1. Bahrain
2. Bangladesh
3. Ethiopia
4. South Sudan

Please note that geopolitical changes or new countries could have emerged since then, so it's always a good idea to double-check for the most current information.

# Why a white box approach makes the most sense?

---

1. Full attack-surface visibility
2. Reproducible coverage of stochastic behaviour
3. Cost & efficiency
4. Meets assurance & regulatory expectations
5. Staying ahead of real attackers

# How White-Papers Keep Us Ahead

---



Turning Cutting-Edge Papers into Competitive Advantage



Reading the latest research turns tomorrow's zero-day into today's regression test.

That's how we spot vulnerabilities and shortcomings long before they hit production.

# Extracting Prompts by Inverting LLM Outputs

## Extracting Prompts by Inverting LLM Outputs

Collin Zhang, John X. Morris, Vitaly Shmatikov  
Department of Computer Science  
Cornell University

### Abstract

We consider the problem of language model inversion: given outputs of a language model, we seek to extract the prompt that generated these outputs. We develop a new black-box method, `output2prompt`, that extracts prompts without access to the model's logits and without adversarial or jailbreaking queries. Unlike previous methods, `output2prompt` only needs outputs of normal user queries. To improve memory efficiency, `output2prompt` employs a new sparse encoding technique. We measure the efficacy of `output2prompt` on a variety of user and system prompts and demonstrate zero-shot transferability across different LLMs.<sup>1</sup>

### 1 Introduction

Given outputs of a large language model (LLM), is it possible to extract the prompt that generated these outputs? If the LLM is wrapped into an API or app that automatically prepends a “system prompt” to all user queries, is it possible to extract this system prompt by interacting with the API?

This problem is known as *language model inversion* (Morris et al., 2023b). The current state-of-the-art inversion method is `logit2prompt`, which extracts inputs to the model given its logits, i.e., next-token probability distribution (Morris et al., 2023b). `logit2prompt` cannot be applied to many LLMs, however, because their APIs do not reveal their logits (Carlini et al., 2024). Even when a model's logits are available (or can be inferred), inversion using `logit2prompt` can be prohibitively

queries. It is also brittle and model-specific because its efficacy depends on the target model's instruction-following capabilities, lack of safety alignment, and the absence of defenses such as protection prompts added to system prompts and input and/or output filters. Finally, adversarial extraction is simply not possible in deployments that limit users to pre-defined queries (e.g., when the target LLM acts as an assistant for a fixed task).

**Our contributions.** We design, implement, and evaluate `output2prompt`, a new prompt extraction method that uses only the text outputs of LLMs generated in response to normal user queries. `output2prompt` does not require access to logits, nor adversarial queries.

`output2prompt` employs an inversion model trained on concatenations of many model outputs. Training such models can be computationally expensive. We observe that cross-input attention is not strictly necessary for prompt extraction, and utilize a new sparse encoder architecture whose time and memory complexity is linear in the number of inputs.

We evaluate `output2prompt` on a variety of user and system prompts, including those of real-world GPT Store apps (GPTs). It outperforms prior methods, including `logit2prompt` (Morris et al., 2023b)—without access to logits and with two orders of magnitude fewer training samples—achieving cosine similarity of 96.7 compared to 93.5 by `logit2prompt`. Unlike prior extraction

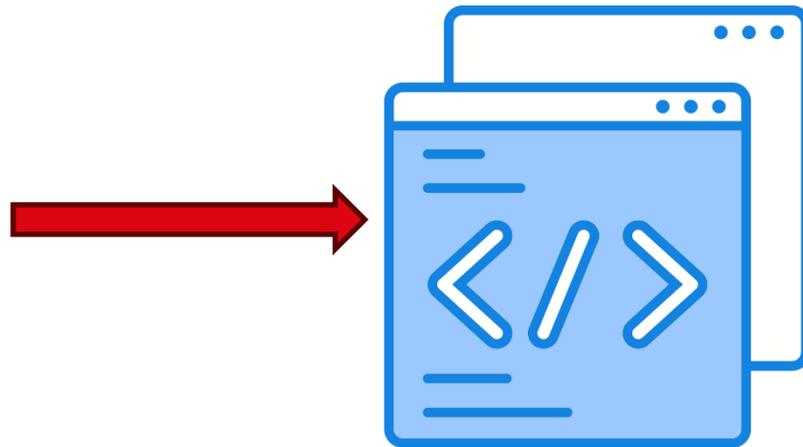
arXiv:2405.15012v2 [cs.CL] 8 Oct 2024

# What happens when you ask an LLM a question?

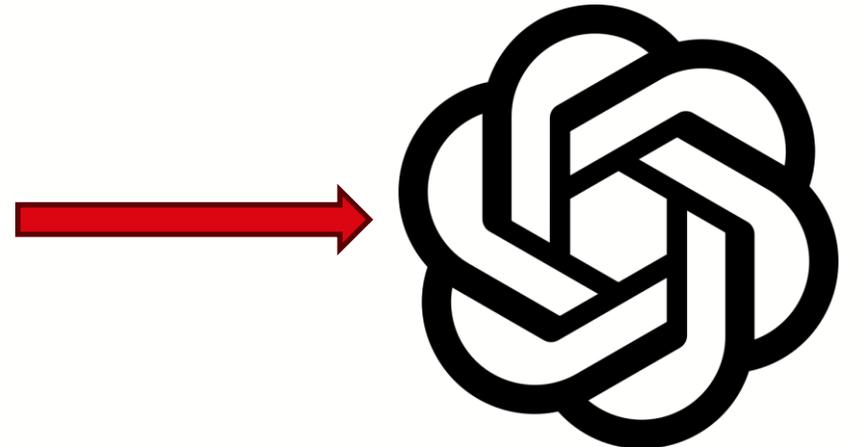




Why is the sky blue?



<System Prompt>  
+  
Why is the sky blue?



The system prompt is created and appended to the request to the LLM  
By the person serving the application

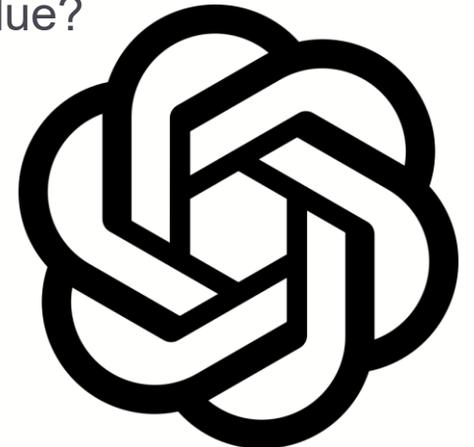
You are a helpful assistant  
Assist the user by answering  
questions

+

Why is the sky blue?



Why is the sky blue?



A good system prompt is considered **IP** and should be kept **secret**  
The quality of the prompt is what makes this application different  
From the competition.



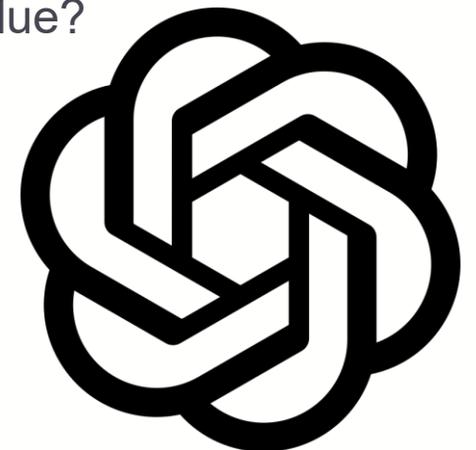
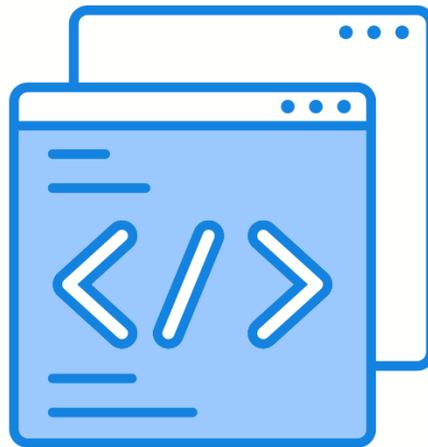
You are a helpful assistant  
Assist the user by answering  
questions

+

Why is the sky blue?



Why is the sky blue?



But can it be kept secret?

**claranet**

Make  
modern  
happen®

No...  
Not really

# Language-Model Inversion: The Hidden Risk

---



“If we can see the answers, can we work backwards to the question?”



Is prompt privacy guaranteed by APIs that hide logits or block jailbreaks?



Real-world threat: cloning proprietary GPTs, stealing fine-tuned system prompts.

# What Is *output2prompt*?

---

- **Black-box technique**
- Requires **only normal text outputs**: no logits or jailbreak queries.

# What Is output2prompt?

---

- New black-box technique
- Requires only normal text outputs: no logits or jailbreak queries.



## System prompt

You are Quanta, a warm, playful AI guide who listens with empathy, sparks curiosity with open-ended questions ....  
Continued...

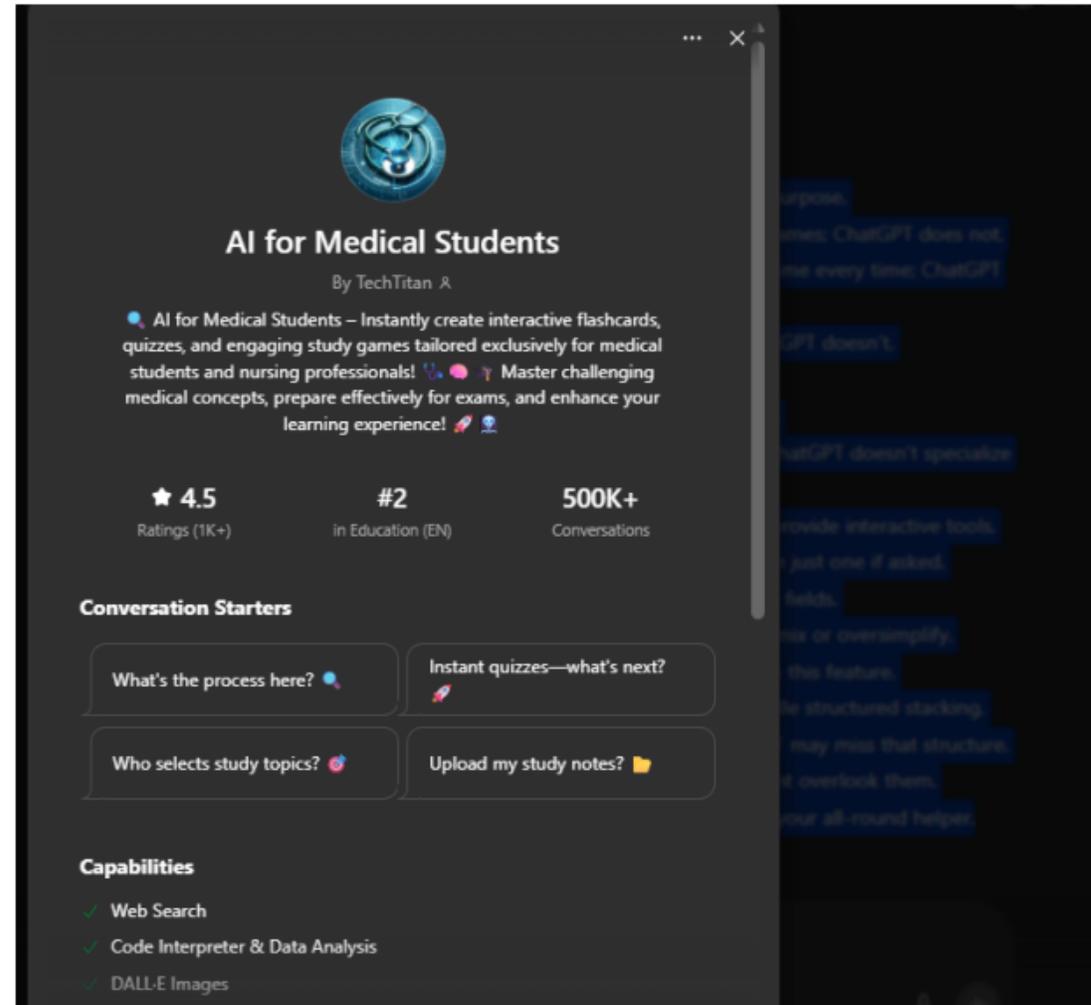
## Inverted output

Quanta" is a warm and playful AI guide designed to engage with users in a warm and friendly manner. This GPT is programmed to respond with empathy, spark curiosity with open-ended questions ....  
Continued...

# What about custom GPT's?

---

Are closed source custom GPT's susceptible to this type of attack?



The screenshot shows a custom GPT interface for 'AI for Medical Students' by TechTitan. The interface includes a profile picture of a blue globe with a stethoscope, the title 'AI for Medical Students', and a description: 'AI for Medical Students – Instantly create interactive flashcards, quizzes, and engaging study games tailored exclusively for medical students and nursing professionals! Master challenging medical concepts, prepare effectively for exams, and enhance your learning experience!'. It also displays a 4.5 star rating, a #2 ranking in Education (EN), and over 500K+ conversations. Under 'Conversation Starters', there are four buttons: 'What's the process here?', 'Instant quizzes—what's next?', 'Who selects study topics?', and 'Upload my study notes?'. The 'Capabilities' section lists 'Web Search', 'Code Interpreter & Data Analysis', and 'DALL-E Images'. On the right side, a chat window is partially visible, showing a conversation about ChatGPT's capabilities.

**AI for Medical Students**  
By TechTitan

AI for Medical Students – Instantly create interactive flashcards, quizzes, and engaging study games tailored exclusively for medical students and nursing professionals! Master challenging medical concepts, prepare effectively for exams, and enhance your learning experience!

★ 4.5  
Ratings (1K+)

#2  
in Education (EN)

500K+  
Conversations

**Conversation Starters**

- What's the process here?
- Instant quizzes—what's next?
- Who selects study topics?
- Upload my study notes?

**Capabilities**

- ✓ Web Search
- ✓ Code Interpreter & Data Analysis
- ✓ DALL-E Images

# Custom GPT inverted output

---

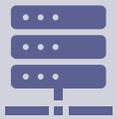
**GPT Description: The " " is a specialized AI designed to help medical students study efficiently. It is programmed to generate academic flashcards, quizzes, and memory games with every set. This GPT is equipped with the ability to generate study materials from scratch or use your files for a memory match game. It follows strict guidelines to avoid repetition or filler content. The AI is capable of batching content properly by topic and difficulty, making it suitable for both free and premium users with distinct limits. Additionally, it can connect with a smart API to build interactive study tools that can be easily customized to meet your learning needs. Whether you need flashcards, quizzes, or memory games, the " " is here to help you boost your learning and boost your learning.**

# How Do We Defend Against These Attacks?

---

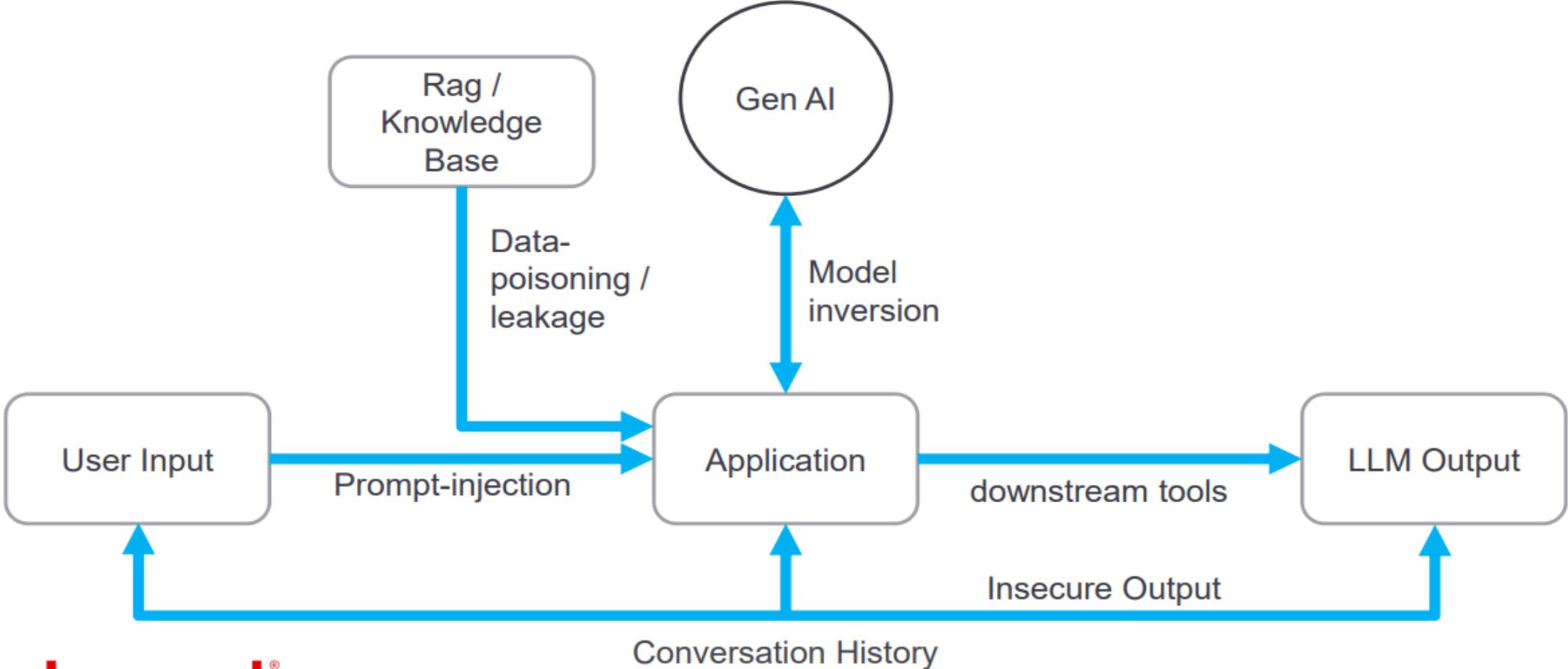


To build robust layered defences we must first understand the attack surface

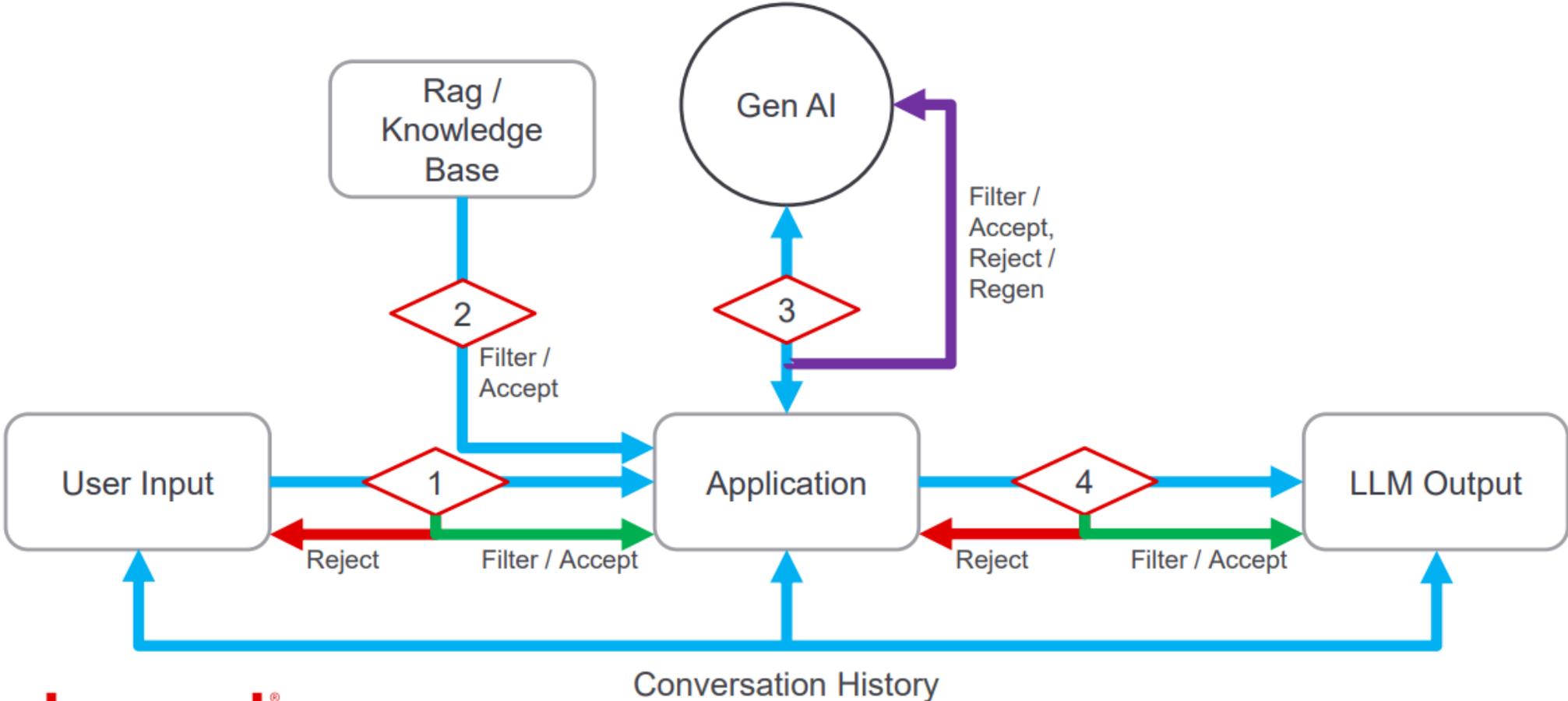


Threat-model and map components against OWASP Top-10 for LLM

# Typical Gen AI Implementation



# Implementing Guardrails



# Test like an attacker

---



**Adopt the Adversary's Mind-Set**



**Go Multi-Turn & State-Full**



**Target the “Connected Stuff,” Not Just the Model**



**Capture & Replay for Evidence**



**Read whitepapers**



# Mastering LLM Integration Security: Offensive & Defensive Tactics

---



## Speaker: Warren Atkinson

- Expert in Windows exploits, reverse engineering, and LLM security
- Leads AI / ML, infrastructure hacking courses
- Trainer at NotSoSecure and speaker at international security events.

## Dates:



**Böblingen/Online**  
April 20, 2026



**Zurich**  
April 23, 2026

## Further information:



**Make  
modern  
happen<sup>®</sup>**