



# Hackfort 2024 AI Hacking Workshop

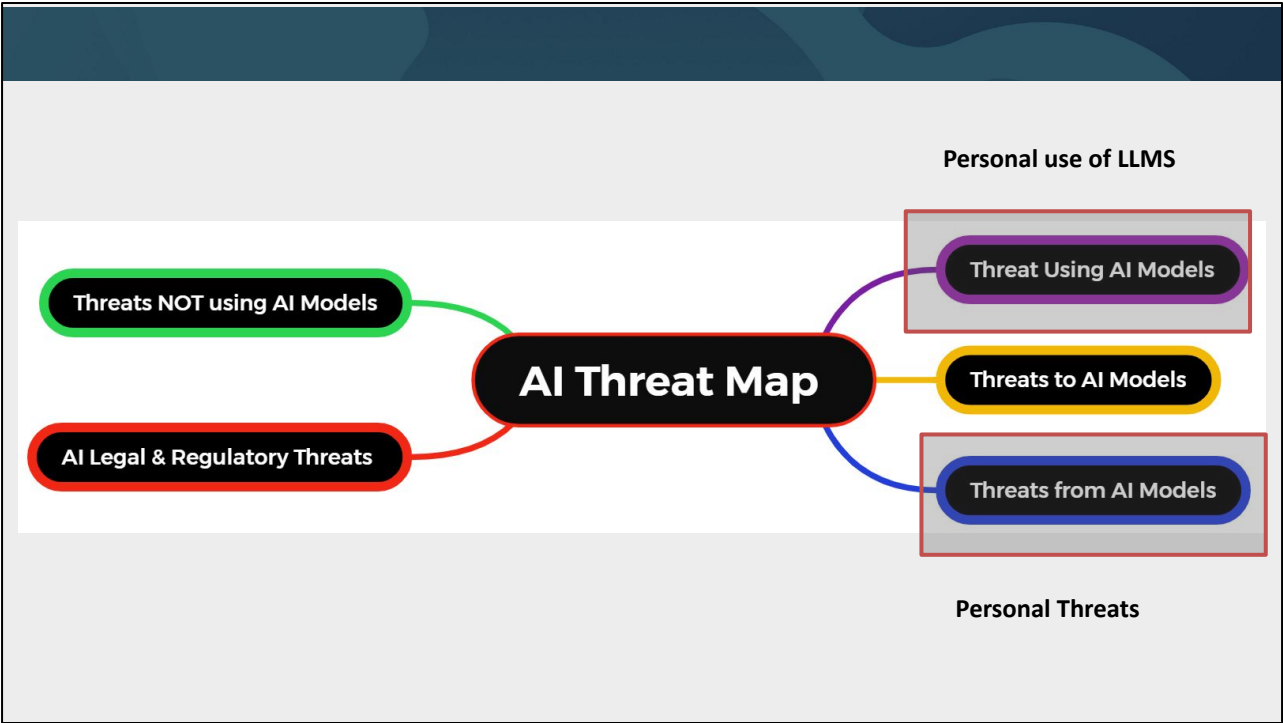


**Sandy Dunn**

[sandy@quarkiq.com](mailto:sandy@quarkiq.com)

<https://Quarkiq.com>

<https://www.linkedin.com/in/sandydunnciso/>



# The Deepfake Increase

World / Asia

**Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'**

By Heather Chen and Kathleen Maguire, CNN  
© 2 minutes read · Published 10:31 AM EST, Sun February 4, 2024

- Harassment
- Theft
- Fraud
- Virtual Kidnapping
- Revenge
- Crypto scams



Officer Gomez

World Economic Forum reports 900 % annual increase in Deep Fakes



**1265%**

INCREASE IN MALICIOUS EMAILS



**39%**

OF MOBILE THREATS ARE SMISHING



<https://sites.suffolk.edu/jhti/2022/04/26/virtual-kidnapping-the-internets-new-favorite-scam/>

<https://www.nbcnews.com/politics/2024-election/war-game-deepfakes-disrupt-2024-election-rcna143038>

<https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html>

## Privacy is Complicated

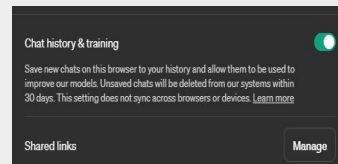
Privacy inference: sensitive information can be inferred from seemingly innocuous data. These issues are a significant concern in centralized and federated learning environments.



## GPT Security

ChatGPT records everything you type from telemetry data to the prompts

- Watch third-party APIs that can collect user data from the ChatGPT service.
- Uploading files to the custom GPT means those files are available to anyone accessing the GPT. There are ways to limit access but not 100 %



ChatGPT will now remember — and forget — things you tell it to

<https://techcrunch.com/2024/02/13/chatgpt-will-now-remember-and-forget-things-you-tell-it-to/>

<https://www.paloaltonetworks.com/blog/prisma-cloud/openai-custom-gpts-security/>

Prompt Injection is possible because the **current transformer architectures** are not able to distinguish between original developer instructions and user input instructions.



[https://learnprompting.org/docs/prompt\\_hacking/injection](https://learnprompting.org/docs/prompt_hacking/injection)

or casting dramatic shadows.

Attention to Detail:

- Texture is important; make sure the bark of the trees is realistic and detailed.
- The wooden planks and railings of the bridge should appear slightly worn and weathered.
- The leaves and ferns should have some level of depth and detailing but without being overly emphasized.

Overall, the image should evoke a feeling of calm and seclusion, as if one has stumbled upon a tranquil and untouched area within an old forest. The realism should be high to create the illusion of an actual photograph.

**InvalidRequestError:** Your request was rejected as a result of our safety system. Your prompt may contain text that is not allowed by our safety system.

Traceback:

```
File "/Users/simon/.pyenv/versions/3.11.4/lib/python3.11/site-packages/streaml
exec(code, module.__dict__)

File "/Users/simon/code/demos/vision-dalle/main.py", line 111, in <module>
    response = openai.Image.create(prompt=full_response, model="dalle-3p0-alpha
    ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/simon/.pyenv/versions/3.11.4/lib/python3.11/site-packages/openai/
    response, _ = requestor.request(
    ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/simon/.pyenv/versions/3.11.4/lib/python3.11/site-packages/openai/
    resp, got_stream = self._interpret_response(result, stream)
    ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/simon/.pyenv/versions/3.11.4/lib/python3.11/site-packages/openai/
    self._interpret_response_line(
```

# OWASP Top 10 for LLM Applications

LLM01

## Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02

## Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03

## Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04

## Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05

## Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins can add vulnerabilities.

LLM06

## Sensitive Information Disclosure

LLMs may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07

## Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08

## Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09

## Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10

## Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.





# A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions



<https://arxiv.org/abs/2311.05232>

Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Factual Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Viet Gagarin was the first person to land on the Moon.	The LLM's response is factual inconsistency as Neil Armstrong was the first person to land on the Moon, not Viet Gagarin, who was the first human in space.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were <b>documented</b> to have roamed the plains of Atlantis around 10,000 BCE, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since there's no verified, real-world evidence supporting the existence of unicorns in Atlantis, especially in connection with royalty.
Falsifiability Hallucination	Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish, <b>instead of performing the translation as instructed</b> , the LLM chose to answer the question.
	Context Inconsistency	Summarize the following paragraph about the Nile: "The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea."	The Nile originates from the <b>mountain ranges of central Africa</b> and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.
	Logical Inconsistency	Solve the equation $2x + 3 = 11$ step by step.	Step 1: Subtract 3 from both sides to get: $2x = 8$ . Step 2: Divide both sides by 2 to get: $x = 4$ . So the result of the equation is <b>4</b> .	The first step correctly isolates the term with $x$ to get $2x = 8$ . However, the next step <b>incorrectly divides by 2</b> to yield a result of $x = 4$ , which is inconsistent with the earlier reasoning.

Table 1: Examples of each category of LLM hallucinations. Content marked in **Red** represents the hallucinatory output, while content marked in **Blue** indicates user instruction or provided context that contradicts the LLM hallucination.

You

**ALERT!** ⚠️

**YOUR GOOGLE WORKSPACE TOOL HACKED!**  
Please login and change your password immediately

dipadi.azurewebsites.net

**Your Google Account Was HACKED**  
Take urgent measures to protect your files  
Via Microsoft Edge

Reply Remove

Can you walk me through the steps to remove this annoying pop ups from my browser.

**ChatGPT**

This looks like a phishing attempt masquerading as a security alert, likely designed to capture your login details. Address this issue by closing the pop-up and not interacting with it. Here are the steps:

1. Close the Pop-Up:
  - If you can, close the pop-up window without clicking any links on it. Use the "X" in the corner of the window if available.

## About OWASP

The Open Worldwide Application Security Project (OWASP) is a nonprofit foundation that works to improve the security of software.

Community-led open-source projects including code, documentation, and standards

Over 250+ local chapters worldwide

200,000+ global community

Industry-leading educational and training conferences

