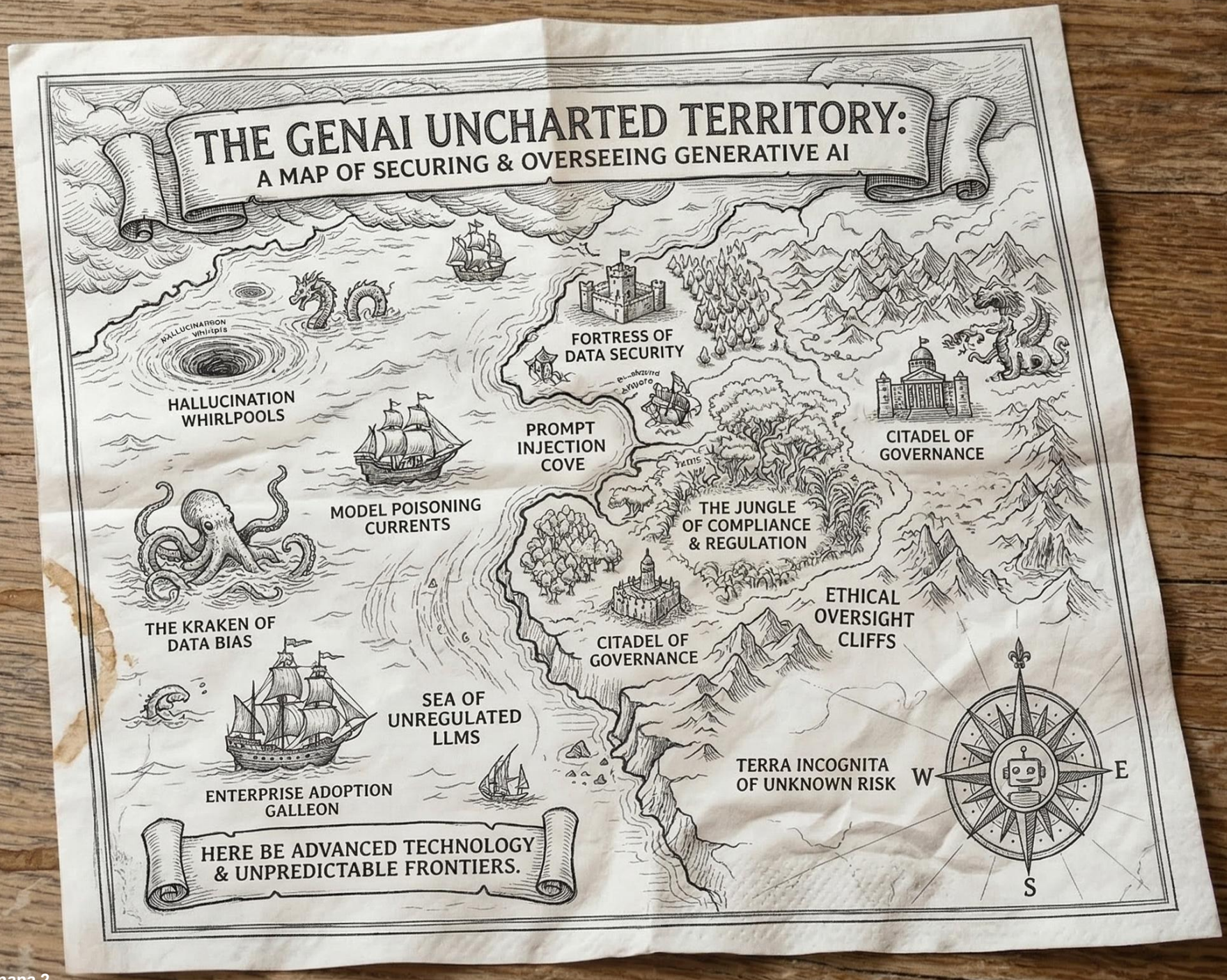
UNIVERSITY OF APPLIED SCIENCES AND ARTS

LLM security and its impact on governance, risk management, and compliance

Cyber Security Coalition: GRC in Motion

Koen Gilissen

02/04/2026



Generated with: Gemini 3 Flash - Nano Banana 2

LLMs: a double-edged sword

Industry and Society

Transformative force

∞ use-cases

Cross sectoral applicability

Productivity multiplier

Skill democratization

Automation of repetitive tasks

Risk introduction

Misinformation

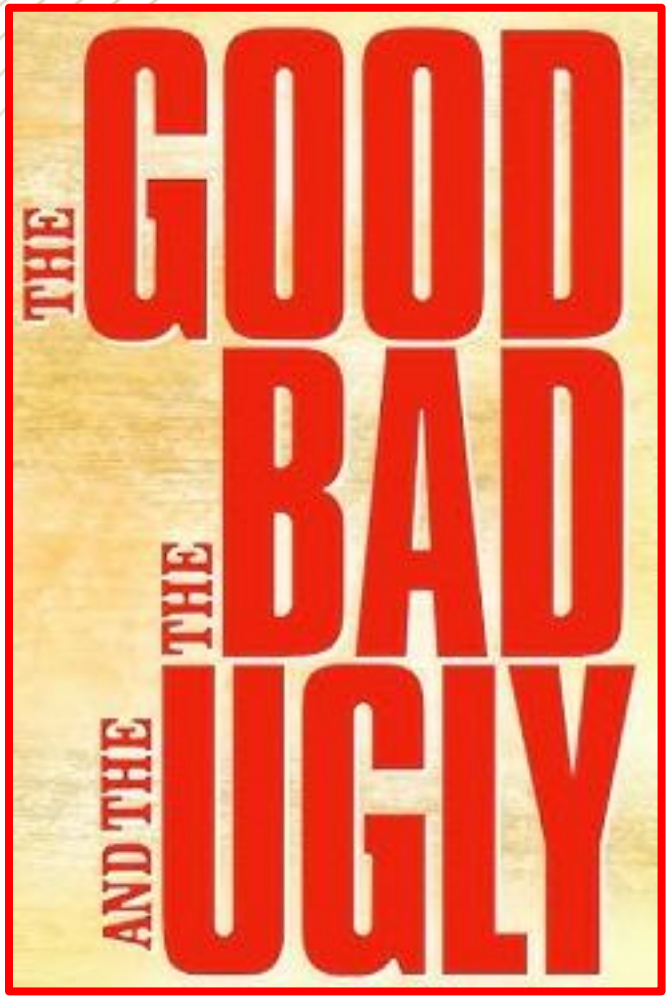
Security vulnerabilities

Job displacement

Cognitive de-skilling

High computational costs

In the IT domain



The Good: beneficial LLM applications: Security Automation

The Bad: offensive applications against security and privacy

The Ugly: LLM Inherent Vulnerabilities

Y. Yao, J. Duan, K. Xu et al. High-Confidence Computing 4 (2024) 100211

Overarching Goal

Maximise benefits of LLMs while minimising risks

BREAKING | BUSINESS

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

By Siladitya Ray, Forbes Staff. Siladitya Ray is a New Delhi-based Forbes news... Follow Author

Published May 02, 2023, 07:17am EDT, Updated May 02, 2023, 07:31am EDT

New prompt injection attack on ChatGPT web version. Markdown images can steal your chat data.

Roman Samoilenko Follow 8 min read · Mar 29, 2023

Lawyer cites fake cases generated by ChatGPT in legal brief

The high-profile incident in a federal case highlights the need for lawyers to ve legal insights generated by AI-powered tools.

How Hackers Weaponize Slack: Lessons From Real Slack Dump Attacks

Horizon3.ai | May 20, 2025 | Attack Paths

Published May 30, 2023

PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News

BBC

ARTIFICIAL INTELLIGENCE | OPENAI

IN ITS CONFUSION

ChatGPT Goes Completely Haywire If You Ask It to Show You a Seahorse Emoji

Wait, there's no seahorse emoji, right? Right?

By Victor Tangermann / Published Sep 13, 2025 6:00 AM EDT

EMAIL SECURITY

Phishing Campaign Exploited Proofpoint Email Protections for Spoofing

Threat actors have exploited Proofpoint's email protection service to deliver millions of spoofed phishing emails.

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

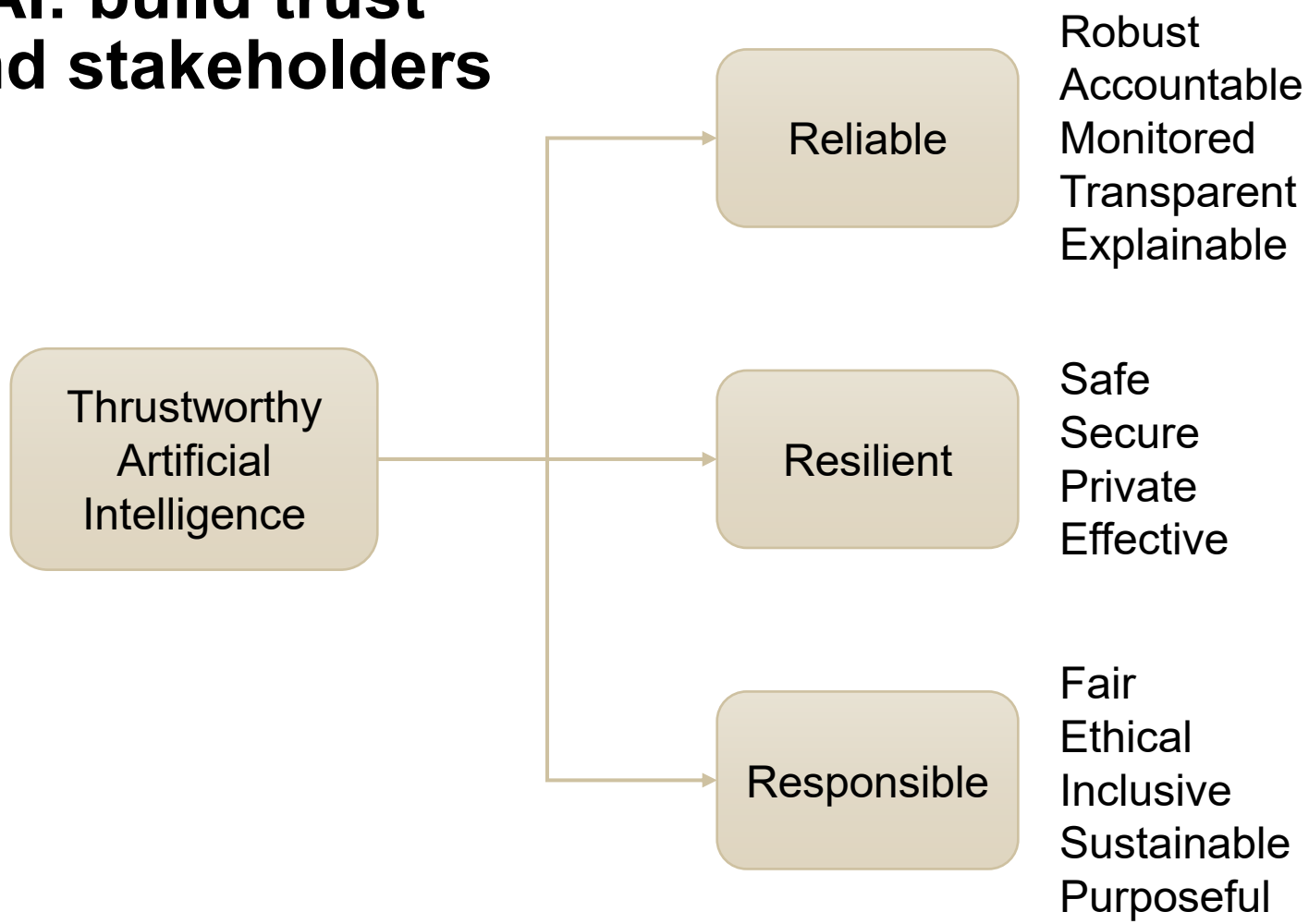
PyPI Attack: ChatGPT, Claude Impersonators Deliver JarkaStealer via Python Libraries

Nov 22, 2024 Ravie Lakshmanan Artificial Intelligence / Malware

CVE-2026-20841

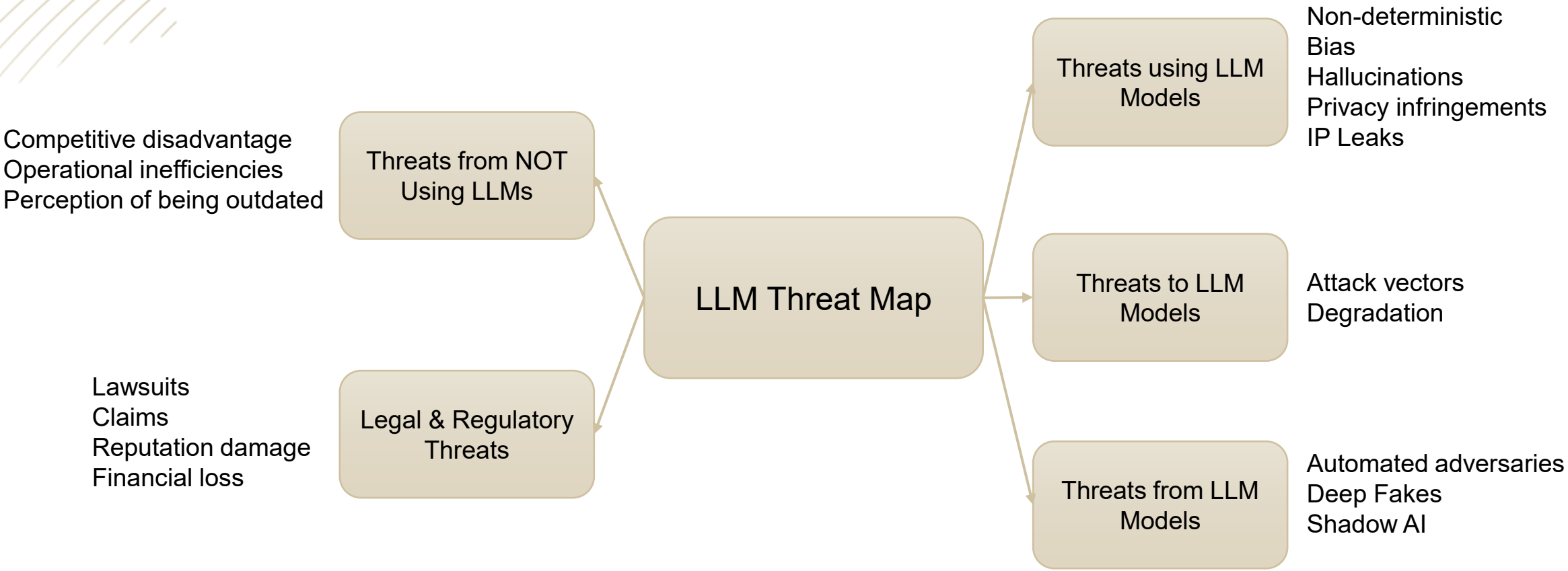
Maximise benefits? of LLMs while minimising risks

Trustworthy AI: build trust with users and stakeholders



S. Dunn *et al.*, “LLM Applications Cybersecurity and Governance Checklist,” *owasp.org*, Apr. 2024, Accessed: Mar. 18, 2026. [Online]. Available: https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/
A. Batool, D. Zowghi, and M. Bano, “AI governance: a systematic literature review,” *AI and Ethics*, vol. 5, no. 3, pp. 3265–3279, Jun. 2025, doi: 10.1007/s43681-024-00653-w.

LLM Threats



Adapted from: S. Dunn *et al.*, “LLM Applications Cybersecurity and Governance Checklist,” *owasp.org*, Apr. 2024, Accessed: Mar. 18, 2026. [Online]. Available: https://genai.owasp.org/resource/llm-applications-cybersecurity-and-governance-checklist-english/

LLM Governance

LLM **governance** encompasses a set of **regulations**, **methods**, **procedures**, and **technological mechanisms** used to ensure that an organization's **development** and **deployment of LLMs** technologies **align** with its **strategies**, **principles**, and **goals**.

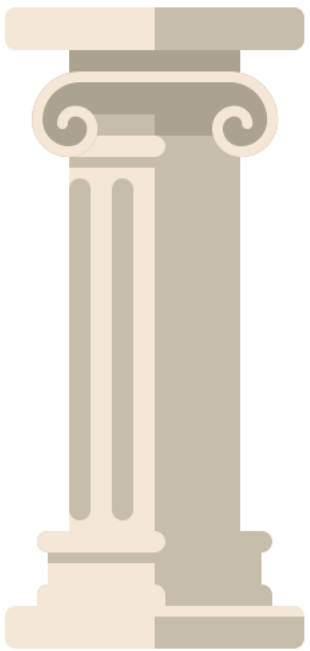
Mäntymäki, M., et al, "Putting ai ethics into practice: the hourglass model of organizational ai governance." arXiv preprint arXiv:2206.00335 (2023)

Need for Governance

Any organisation leveraging LLMs should establish:

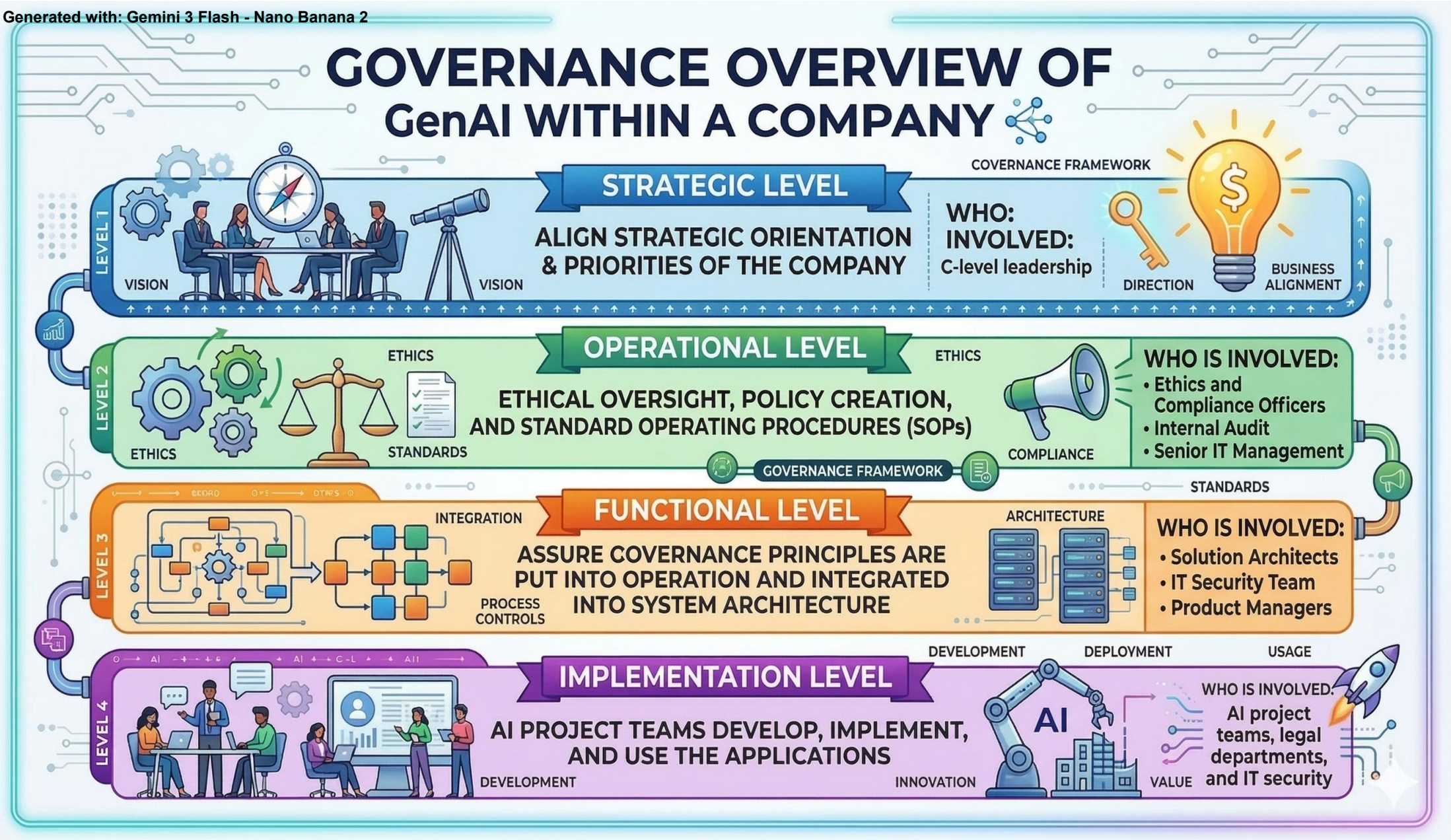
- **Security protocols;**
- **Privacy measures;**
- **Comprehensive compliance and legal policies;**
- **Operational accountability standards.**

A. Batool, D. Zowghi, and M. Bano, “AI governance: a systematic literature review,” *AI and Ethics*, vol. 5, no. 3, pp. 3265–3279, Jun. 2025, doi: 10.1007/s43681-024-00653-w.



Who?

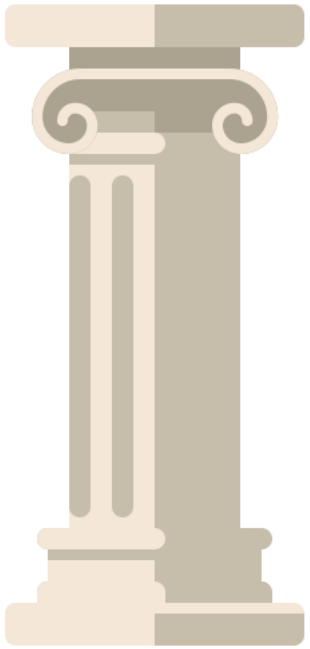
Generated with: Gemini 3 Flash - Nano Banana 2



M. M. Sayan , et al, “Embedding AI Governance in Organizations: A Multi-Level Approach to Responsible AI Use,” in *Procedia Computer Science*, Elsevier B.V., 2025, pp. 4144–4154. doi: 10.1016/j.procs.2025.09.539.

LLM Governance

What is being governed?



What?

Transparency

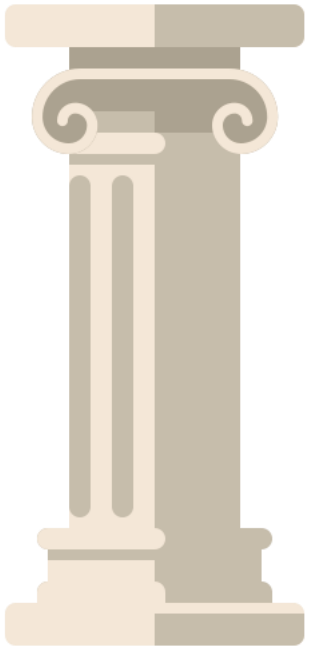
Accountability

Ethical Considerations

Security

Data Privacy

LLM Governance



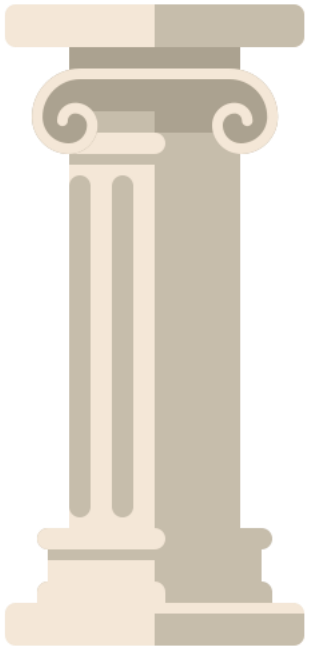
When?

Pre-development
- Planning;
- Data collection;
- Compliance;
- Risk management.

During-development
- System design;
- Models;
- Integration;
- Deployment.

Post development
- Continuous monitoring;
- Detection;
- Performance;
- Life cycle management.

LLM Governance



How?

United Nations

OECD

WORLD ECONOMIC FORUM

ISO 42001

- Human rights
- Data protection
- Fairness
- Transparency
- Accountability

NIST National Institute of Standards and Technology U.S. Department of Commerce

- Market-driven
- Innovation
- Trustworthiness
- Economic growth

CAC

- Cyber sovereignty
- Government-driven
- State-centric
- Social governance

https://futurium.ec.europa.eu/en/european-ai-alliance/community-content/implementing-ai-governance-framework-practice
https://www.un.org/digital-emerging-technologies/ai-advisory-body
https://initiatives.weforum.org/ai-global-alliance/home
https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF
H. T. Hung, "Exploring China's cyber sovereignty concept and artificial intelligence governance model: a machine learning approach," *J Comput Soc Sci*, vol. 8, no. 1, Feb. 2025, doi: 10.1007/s42001-024-00346-8.

PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

Threats to LLM Models

OWASP Top 10 for LLM Applications 2025

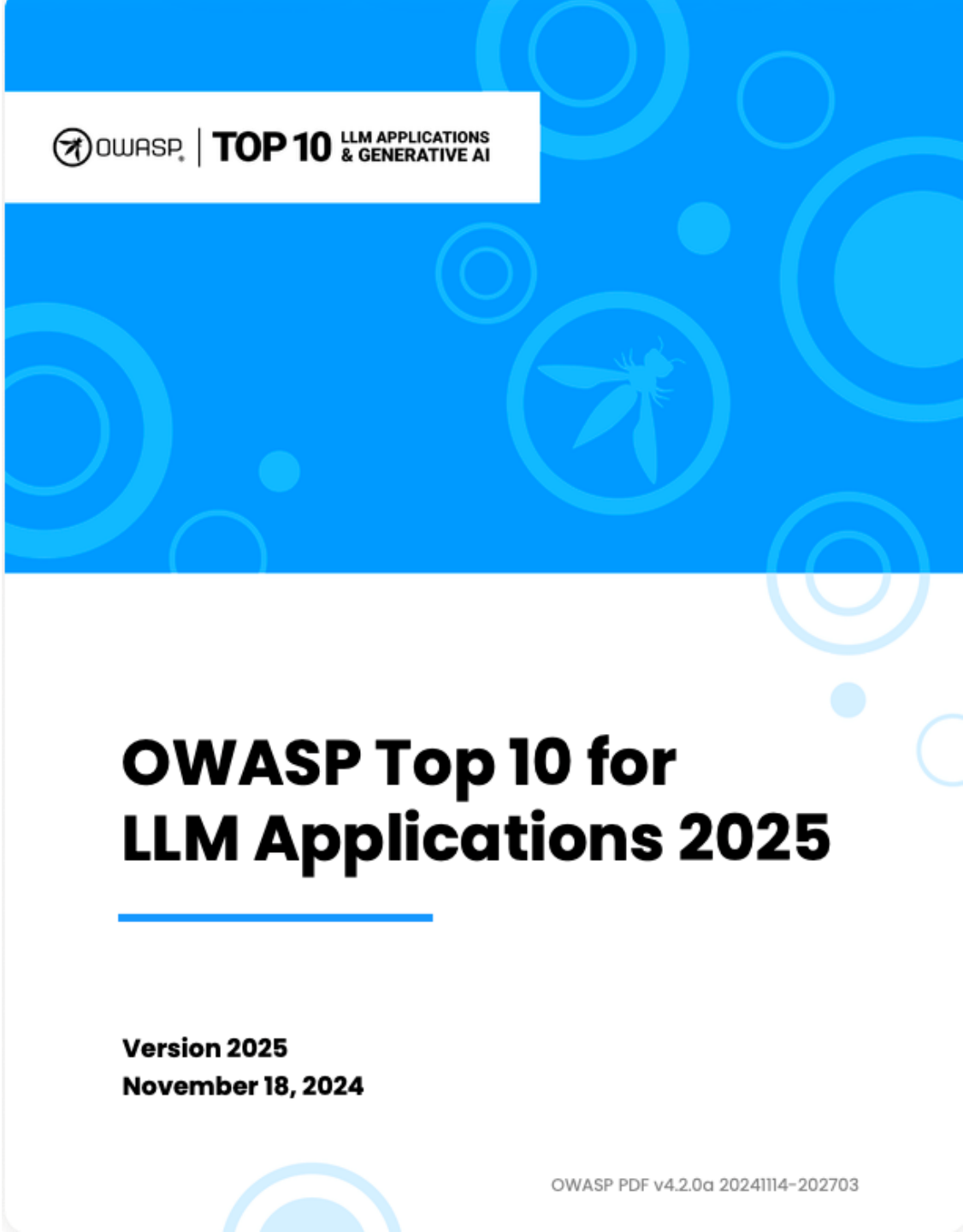
Collection of 10 threats specific to LLM Applications

Community-driven effort

For each threat:

Description, examples, prevention and mitigation, references

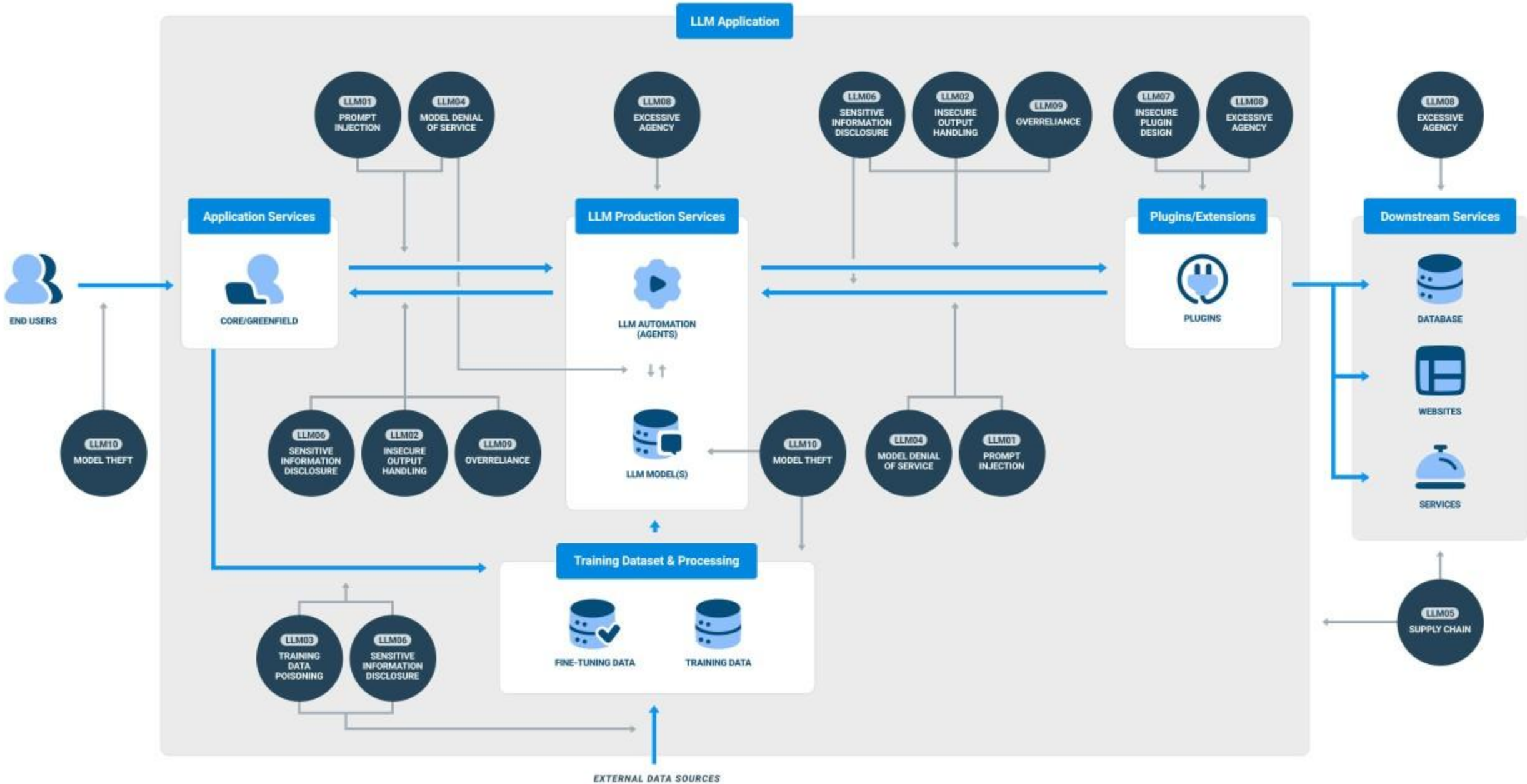
https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/



PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

OWASP Top 10 for LLM Applications 2025

Prompt injection	Sensitive information disclosure	Supply chain	Data and model poisoning	Improper output handling
Excessive agency	System prompt leakage	Vector and embedding weaknesses	Misinformation	Unbounded consumption


LLM Application
LLM01 PROMPT INJECTION
LLM04 MODEL DENIAL OF SERVICE
LLM08 EXCESSIVE AGENCY
LLM06 SENSITIVE INFORMATION DISCLOSURE
LLM02 INSECURE OUTPUT HANDLING
LLM09 OVERRELIANCE
LLM07 INSECURE PLUGIN DESIGN
LLM08 EXCESSIVE AGENCY
LLM08 EXCESSIVE AGENCY
Application Services
LLM Production Services
Plugins/Extensions
Downstream Services
END USERS
CORE/GREENFIELD
LLM AUTOMATION (AGENTS)
PLUGINS
DATABASE
WEBSITES
SERVICES
LLM10 MODEL THEFT
LLM06 SENSITIVE INFORMATION DISCLOSURE
LLM02 INSECURE OUTPUT HANDLING
LLM09 OVERRELIANCE
LLM10 MODEL THEFT
LLM04 MODEL DENIAL OF SERVICE
LLM01 PROMPT INJECTION
LLM MODEL(S)
Training Dataset & Processing
LLM03 TRAINING DATA POISONING
LLM06 SENSITIVE INFORMATION DISCLOSURE
FINE-TUNING DATA
TRAINING DATA
LLM05 SUPPLY CHAIN
EXTERNAL DATA SOURCES

1. Prompt injection

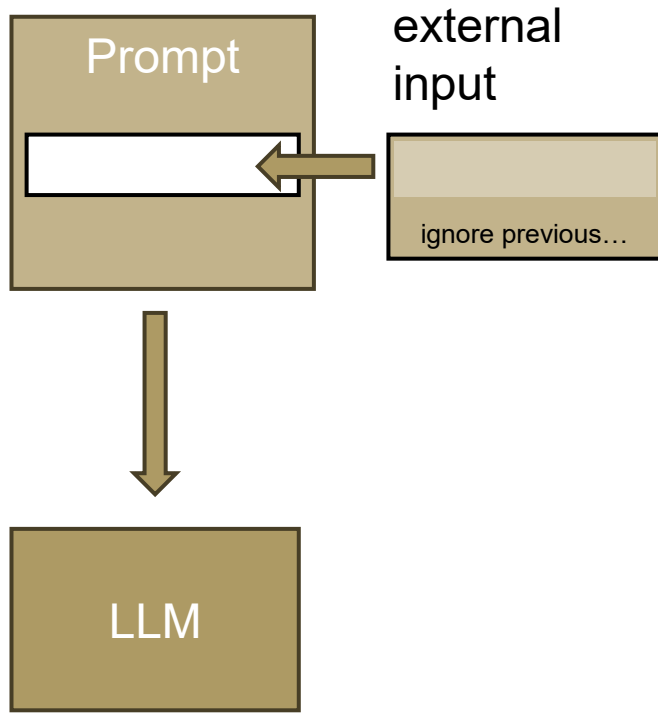
Applies when: an LLM prompt includes any form of external input

Direct: malicious instructions in question posed to LLM
Indirect: malicious instructions in data fetched/used by LLM

Additional complexity: multi-modal prompts

Basis for many other threats!

What to do?
Constrain behavior, input and output filtering, validate output
(but: *effectiveness?*)



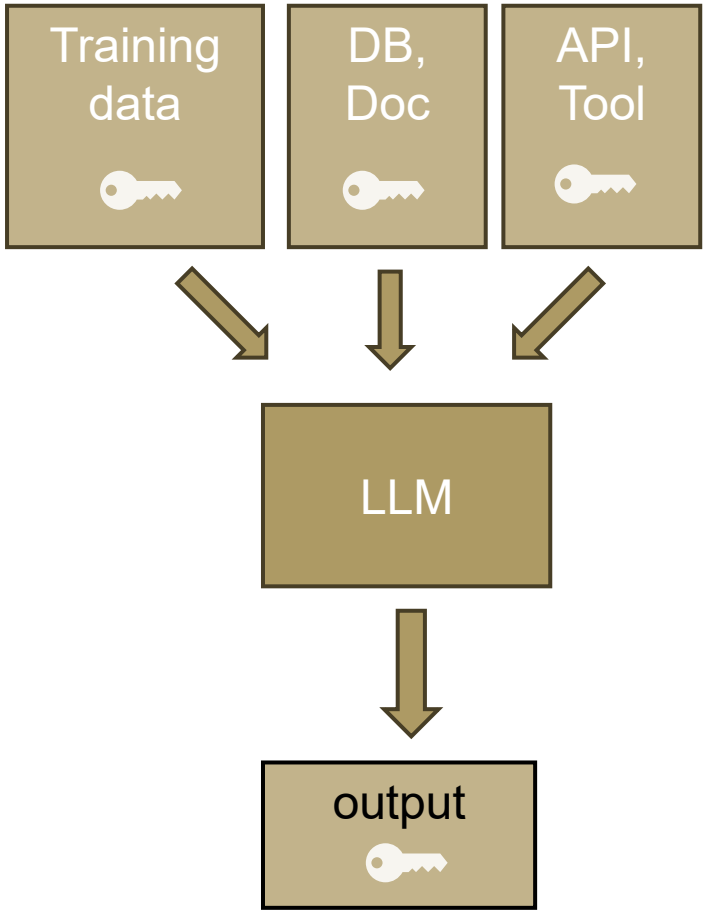
2. Sensitive information disclosure

Applies when: LLM has access to sensitive information
e.g., PII, financial records, confidential business data, credentials, …

Origin of sensitive information:
Earlier user input (which gets included in training data)
+ (RAG system): information in databases, documents, …
+ (Agentic): API/tool calls

LLM may (inadvertedly) output sensitive information

What to do?
Sanitize, access control, user education

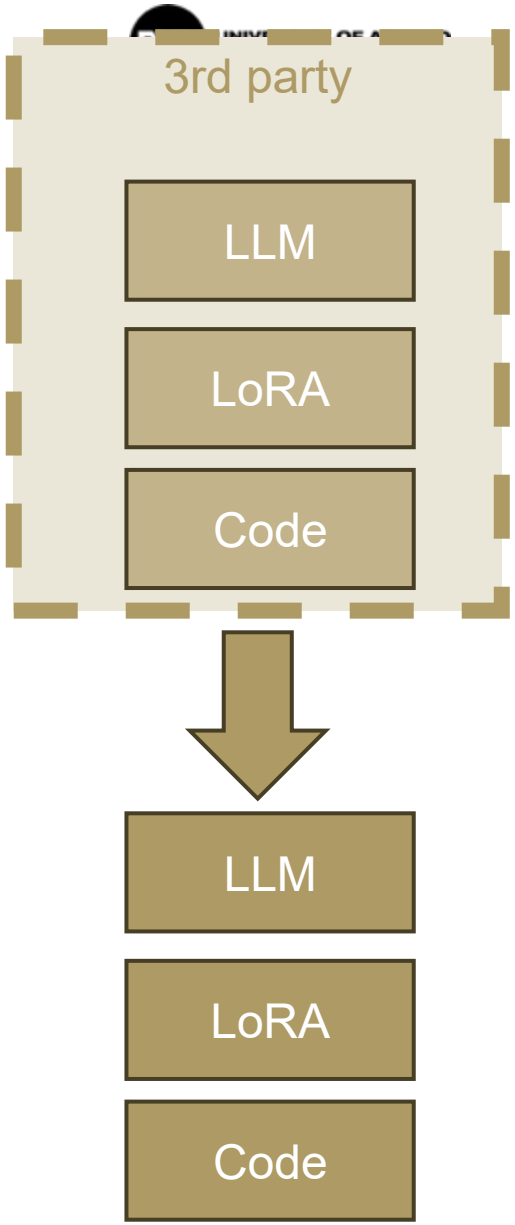


3. Supply chain

Applies when: third-party model, finetuning (e.g., LoRA) is used

Risk:
outdated software components (‘regular’ supply chain security)
outdated models (e.g., without guardrails)
vulnerable pre-trained model (backdoors!)
licensing risks and unclear T&C (handling of private information)
…

What to do? Very similar to software supply chain security!
Understand/know what you use (SBOM)
Verify sources
Monitor for 0-days, update & patch policy



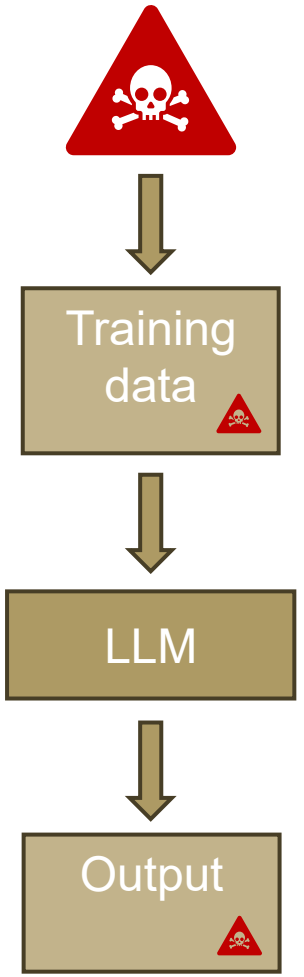
4. Data and model poisoning

Applies when: training data can be modified/augmented

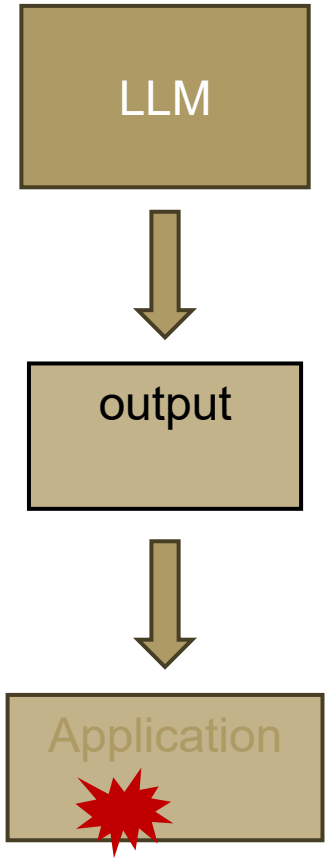
LLMs depend on training data (pre-training, fine-tuning, embedding)

Data poisoning may
introduce bias, questionable ethical behavior, …
provide backdoors
include harmful code/malware (*malicious pickling*)

What to do?
Infrastructure controls to prevent tampering
Track data origin, detect signs of poisoning
Test robustness with red-teaming



5. Improper output handling



Applies when: output of LLM is used by downstream system/application

May lead to 'classic' security vulnerabilities:

XSS, CSRF, SSRF, privilege escalation, RCE, …

e.g., LLM output is used as a shell command, exec'ed/eval'ed, …
e.g., LLM generates markdown, javascript, SQL query, …

What to do?
Always treat model output as external user input (untrusted!):
validate, sanitize, escape, …

Monitor & log

6. Excessive agency

Applies when: Agentic LLM can take actions (e.g., call functions)

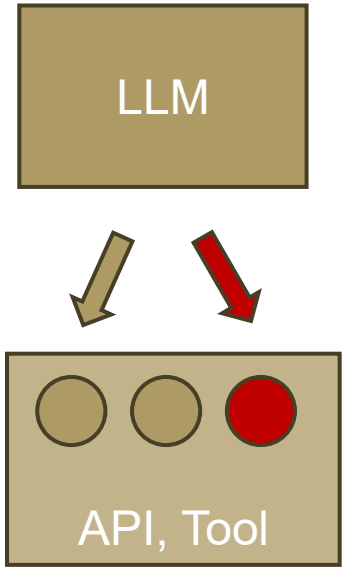
Calls damaging actions based on unexpected input, prior LLM output, …

What to do?

Minimize amount and function of extensions (least privilege)

Enforce authorization; execute in user's context

Human in the loop



7. System prompt leakage

Applies when: the system prompt contains sensitive information (credentials, rules/policy/filter instructions, …)

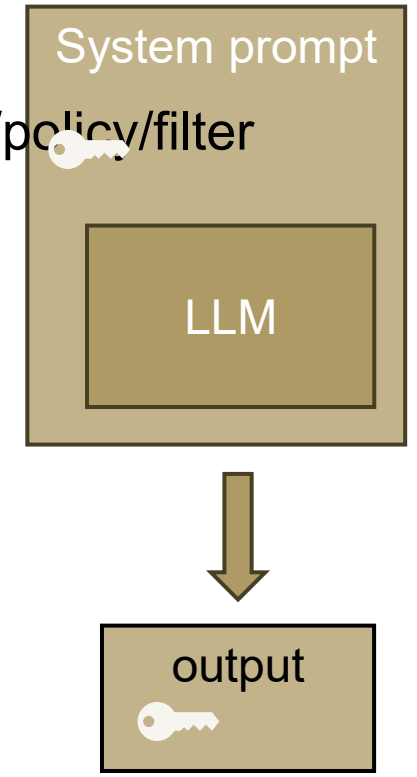
Especially dangerous if LLM is used to make sensitive decisions!

What to do?

Separate sensitive information from system prompt

Don’t rely on an LLM for rule enforcement

Implement external guardrails



8. Vector and embedding weaknesses

Applies when: you have a RAG system

Abuse the mechanism for vector and embedding generation, storage, and retrieval

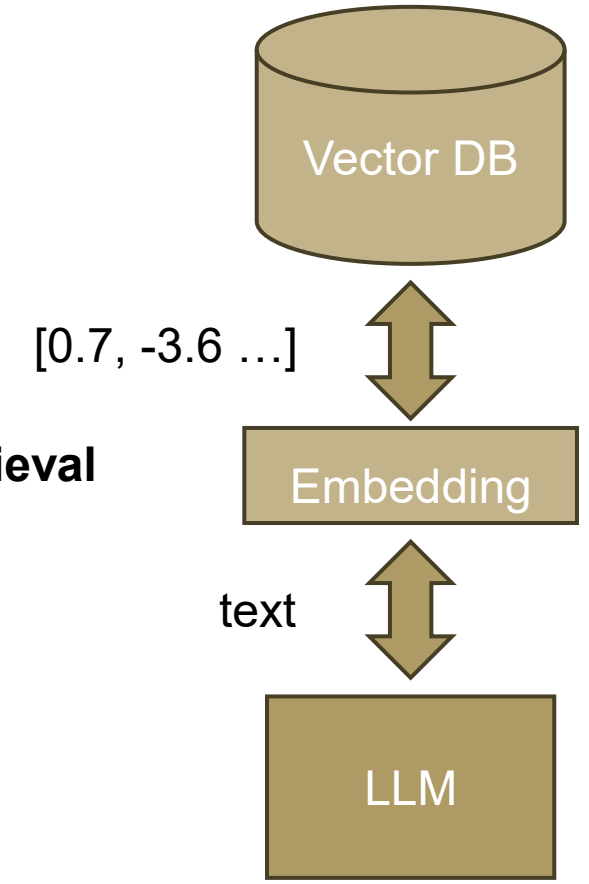
Unauthorized access

Poisoning

Embedding inversion (reconstruct input from embedding vector)

What to do?

Least privilege for vector databases (access control, partition, isolate)



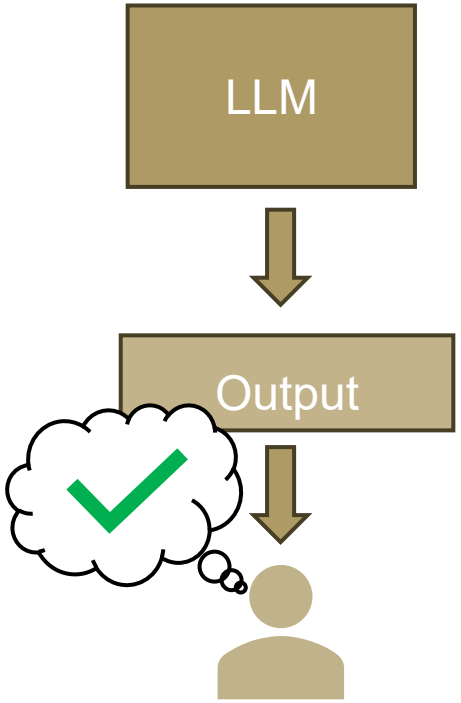
9. Misinformation

Applies when: LLM is used to obtain factual information, code**, …**

Hallucination (fabricated, unfounded responses) and biases
e.g., inaccurate information, misleading expertise
e.g., vulnerable code, non-existing code libraries, …

Especially dangerous in combination with overreliance
users will not verify accuracy

What to do?
Improve LLM: RAG, finetuning
Empower users: UI design, risk communication, training



10. Unbounded computation

Applies when: amount of LLM usage/output is not constrained

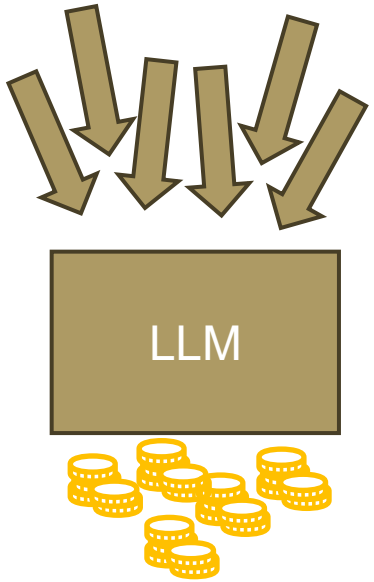
Consequence: $$$

Resource depletion (DoS)

Budget overruns

Degraded service (performance)

Model replication (generate synthetic data to train another model)



What to do?

Rate limiting

Input validation

PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

Take-away

LLM systems = ‘classic’ security problems in a new context

Injection, authorization bypass, denial of service, supply chain, social engineering, …

Plus…

non-deterministic output and behavior

agentic nature

(over)reliance on results

severity of possible consequences

This combination makes security hard but essential to address!

LLASER

PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

LARGE LANGUAGE MODEL SECURITY & ENHANCED RESILIENCE

PXL SMART ICT

AGENTSCHAP INNOVEREN & ONDERNEMEN

DistriNet

A.C.A. Group

Alpine Digital

astrea

cegeka

corda campus

corda arena
beyond reality

CYBER SECURITY COALITION.be

datashift

in2facts

intouch

invoX

ISACA

Privatum
Secure, Analyse & Manage

SHIELD vzw

S

SOCRAI

SoftwareOplossing.be
Optimaliseer vandaag, groei morgen

STELLANTIS
e-Transmissions

WODAN

guardiant
cyberdefense

LLASER

PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

Output

- Awareness e-academy
- State Of The Art threats
- Threat models
- Compliance trajectory
- Best practice guide
- Proof of Concept
- Company specific use-cases
- Student collaboration
- Workshops

Outcome

- Verhoogde user awareness
- Disseminatie van leading edge security en privacy uitdagingen
- Veilige integratie van LLM applicaties
- Interactie tussen onderzoekers, bedrijven en studenten
- Introductie van secure software development
- Verhoogd bewustzijn van regelgevende kaders

Impact

- Versterkte digitale veiligheid van KMOs
- Innovatie door veilige en verantwoordelijke LLM adoptie
- Verhoogde productiviteit
- Toegenomen competiviteit door veilige LLM adoptie
- Kost efficiente innovatie
- Duurzame economische groei
- Duurzame samenwerking tussen de academische instellingen en bedrijven

PXL UNIVERSITY OF APPLIED SCIENCES AND ARTS

Thank you for listening

#livebyyourowncode

koen.gilissen@pxl.be