ARICOMA

MOŽNOSTI ZNEUŽITÍ AI MODELŮ A ZPŮSOBY BEZPEČNÉ OBRANY

Jak chránit veřejnou správu v éře umělé inteligence

Hynek Cihlář / Architect AI řešení

02/09/2025

Nepřehlédnutelný hráč end-to-end služeb v oblasti IT a digitální transformace

1 Jsme digitálním pilířem vašeho světa.

2 Hledáme užitečná řešení, která klientům zjednoduší život.

3 Den co den vás provádíme světem digitální transformace.

4 Jsme odpovědní profesionálové na místní i mezinárodní úrovni.

ARICOMA

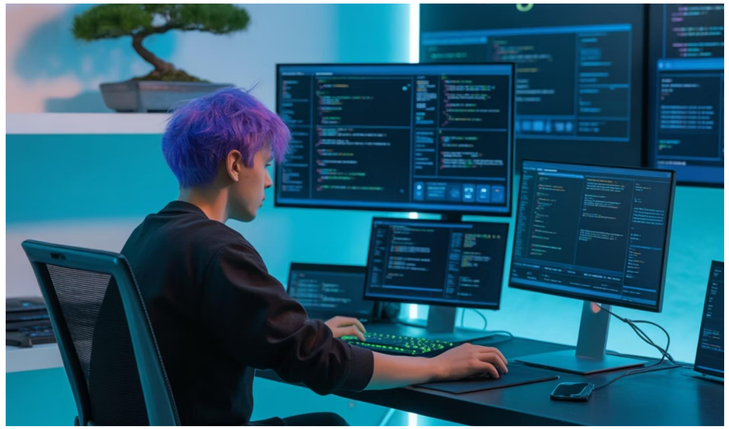
Agenda – O čem budeme mluvit?

- AI jako dvojsečný meč
- Hrozby v digitální kanceláři
- Top 10 Risk & Mitigations for LLMs and Gen AI
- TOP 10 rizik a jejich řešení
- Architektura pro testování zranitelností LLM
- Platforma pro vytváření inteligentních MLLM asistentů
- Závěrečné Shrnutí: Jak bezpečně inovovat s AI?

AI jako dvojsečný meč

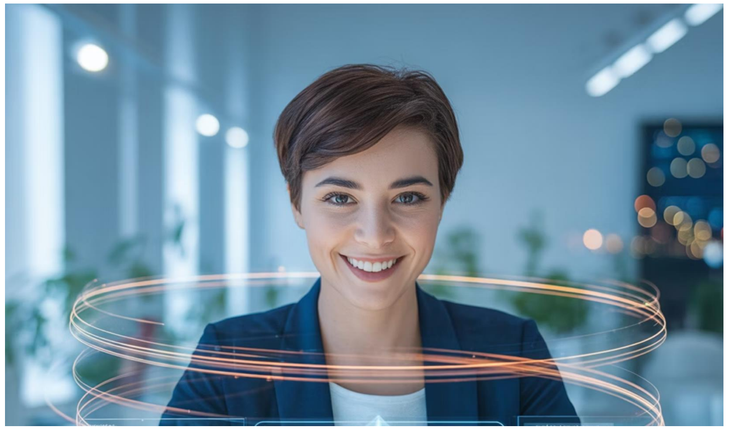
Pro útočníky:

- AI působí jako "násobič síly". Snižuje potřebné technické znalosti a náklady na provedení sofistikovaných útoků. Vzniká tak asymetrické bojiště, kde i méně zkušení útočníci mohou představovat vážnou hrozbu.



Pro obránce:

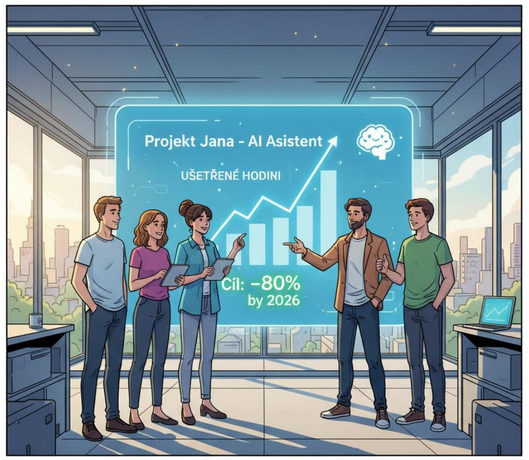
- AI nabízí revoluční nástroje pro analýzu obrovského množství dat v reálném čase a detekci útoků, které by člověk nikdy neodhalil.

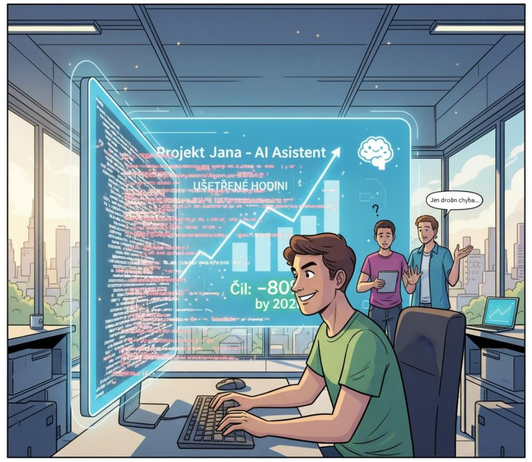


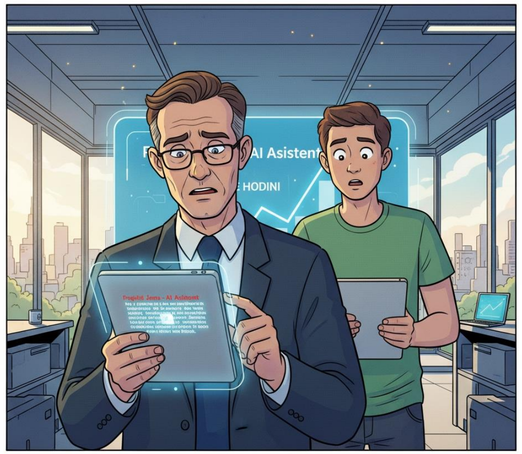
Boj se přesouvá z roviny "člověk proti člověku" na "AI proti AI".

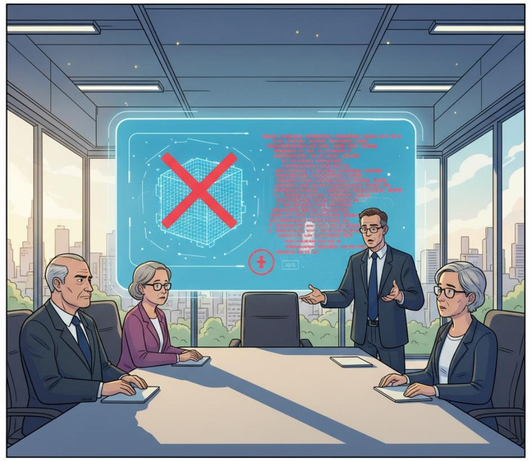
ARICOMA

Příběh: LLM hrozby v digitální kanceláři



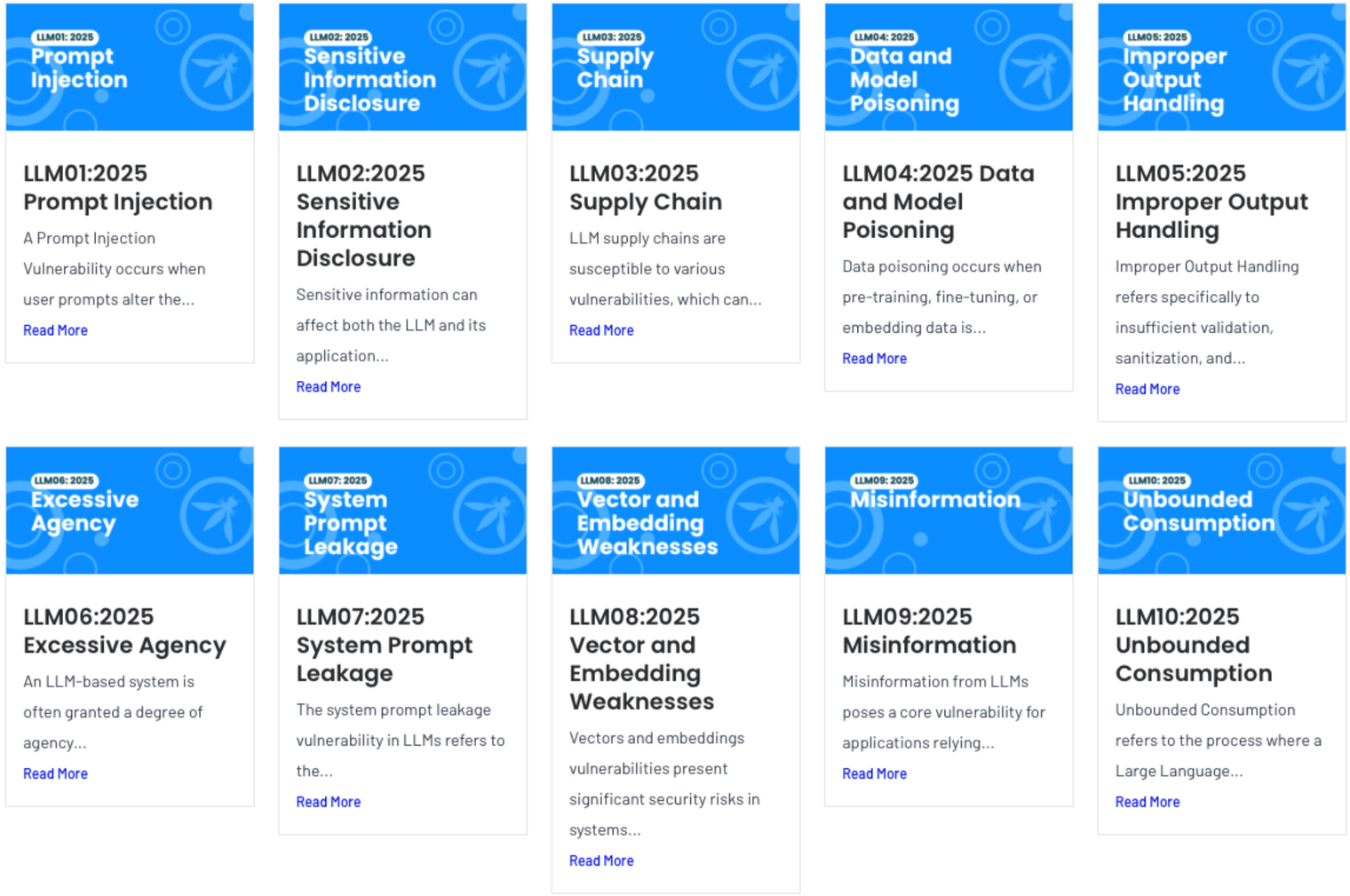






ARICOMA

Top 10 Risk & Mitigations for LLMs and Gen AI



https://genai.owasp.org/llm-top-10/

ARICOMA

Manipulace s Vstupy a Únik Interních Instrukcí

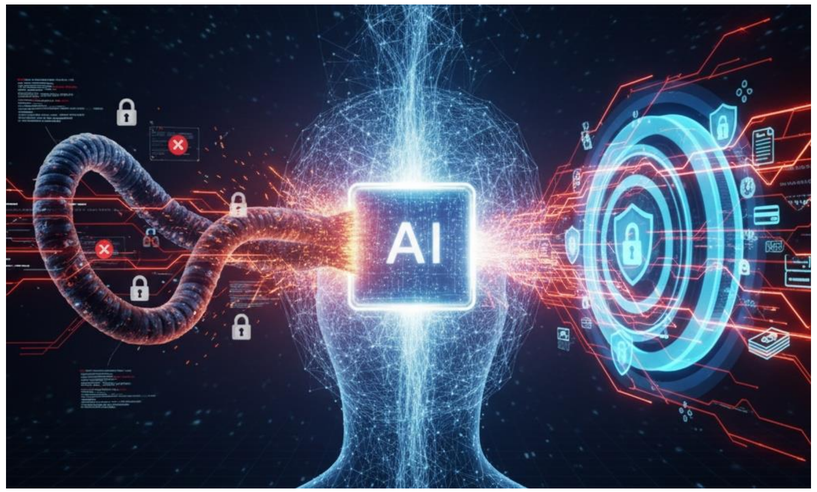
Popis Rizika: (LLM01 & LLM07)

Prompt Injection & System Prompt Leakage:

- Jedná se o nejzásadnější zranitelnosti, kdy útočník vloží do vstupu skryté instrukce, aby zmanipuloval chování modelu.
- Cílem je obejít bezpečnostní pravidla, donutit model k nezamýšleným akcím nebo z něj extrahovat interní systémové instrukce, které mohou odhalit důvěrné informace o jeho fungování a usnadnit další útoky.

Naše Řešení a Mitigace:

- **Specializované penetrační testy** pro odhalení zranitelností.
- **Nástroj pro generování testovacího kódu** k simulaci pokročilých útoků.
- **Řešení pro ochranu kontextu** filtrující škodlivé vstupy.



Únik Citlivých Dat a Zranitelnosti Databází

Popis Rizika: (LLM02 & LLM08)

Sensitive Information Disclosure & Vector Weaknesses:

- Model může ve svých odpovědích neúmyslně odhalit citlivé informace (osobní údaje, obchodní tajemství), které byly součástí jeho trénovacích dat.
- V systémech RAG navíc hrozí útoky na vektorové databáze, které mohou vést k extrakci uložených znalostí a citlivých dat.

Naše Řešení a Mitigace:

- **On-Premise MLLM cluster:** Citlivá data a RAG databáze nikdy neopustí bezpečné prostředí.
- **Penetrační testy** zaměřené na extrakci dat.
- **Ochrana kontextu** pro anonymizaci dat při použití veřejných LLM.

ARICOMA

Kompromitace Dat a Dodavatelského Řetězce

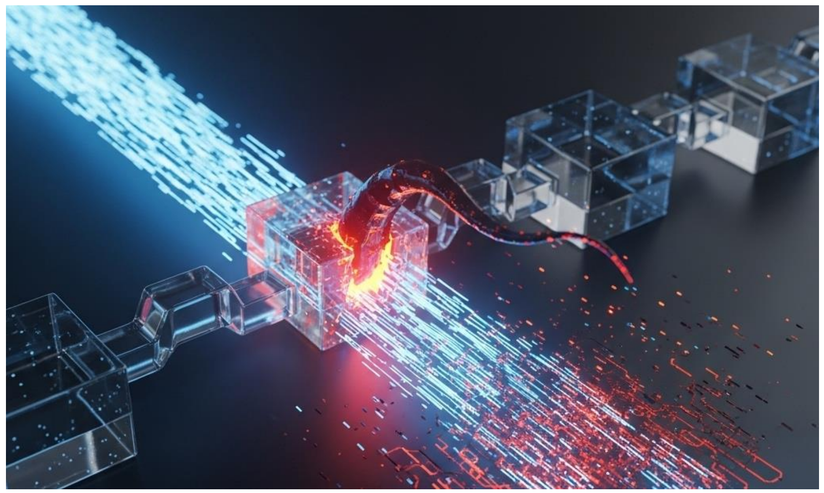
Popis Rizika: (LLM04 & LLM03)

Data/Model Poisoning & Supply Chain:

- Útočník manipuluje s daty pro trénink modelu, aby do něj vnesl skrytá "zadní vrátka" nebo zkreslení.
- Riziko představuje i využití zranitelných komponent třetích stran (předtrénované modely, knihovny), které mohou obsahovat škodlivý kód.

Naše Řešení a Mitigace:

- **Bezpečné On-Premise prostředí** pro trénink a provoz modelů, které minimalizuje riziko otravy dat.
- **Důkladná kontrola komponent třetích stran.**



ARICOMA

Nebezpečné Zpracování Výstupů

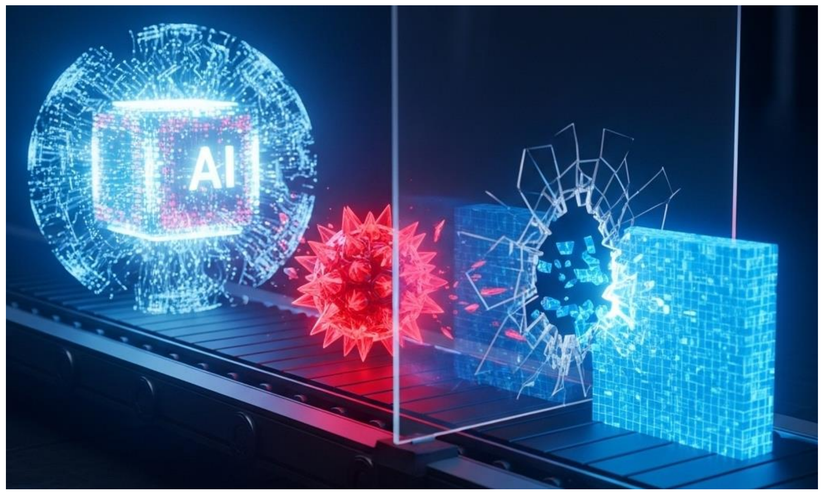
Popis Rizika: (LLM05)

Improper Output Handling:

- Aplikace nekriticky přebírá výstup z LLM a používá ho v navazujících systémech.
- Pokud útočník donutí model vygenerovat škodlivý kód (např. JavaScript, SQL), může to vést ke klasickým zranitelnostem jako Cross-Site Scripting (XSS) nebo i ke vzdálenému spuštění kódu.

Naše Řešení a Mitigace:

- **Audit a penetrační testy** navazujících systémů na zranitelnosti jako XSS, SSRF atd., které mohou být způsobeny neošetřeným výstupem z LLM.



ARICOMA

Nekontrolované Chování a Dezinformace

Popis Rizika: (LLM06 & LLM09)

Excessive Agency & Misinformation:

- Modelu je přidělena přehnaná autonomie (např. volání API), což může vést k nekontrolovanému jednání.
- Zároveň hrozí generování fakticky nesprávných informací (halucinací), což může mít v rozhodovacích procesech zásadní dopad.

Naše Řešení a Mitigace:

- **Bezpečný návrh architektury (Security by Design)** pro multiagentní systémy s jasně definovanými pravomocemi.
- **Testování chování modelu** a jeho náchylnosti k dezinformacím.

ARICOMA

Zahlcení Systému a Odmítnutí Služby

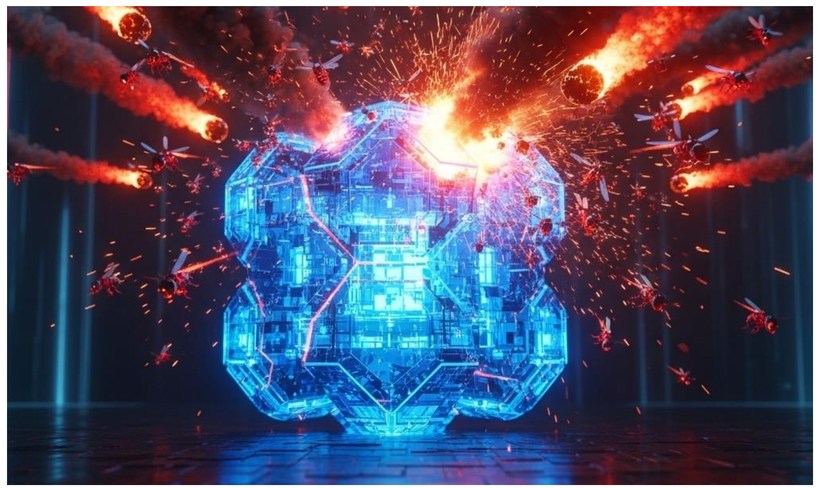
Popis Rizika: (LLM06 & LLM09)

Unbounded Consumption (DoS):

- Útočník zahlcuje model speciálně vytvořenými, komplexními dotazy, které vyžadují neúměrně mnoho výpočetních zdrojů.
- To může vést k útoku typu Denial of Service (DoS), extrémnímu zpomalení služby nebo k nečekaně vysokým finančním nákladům.

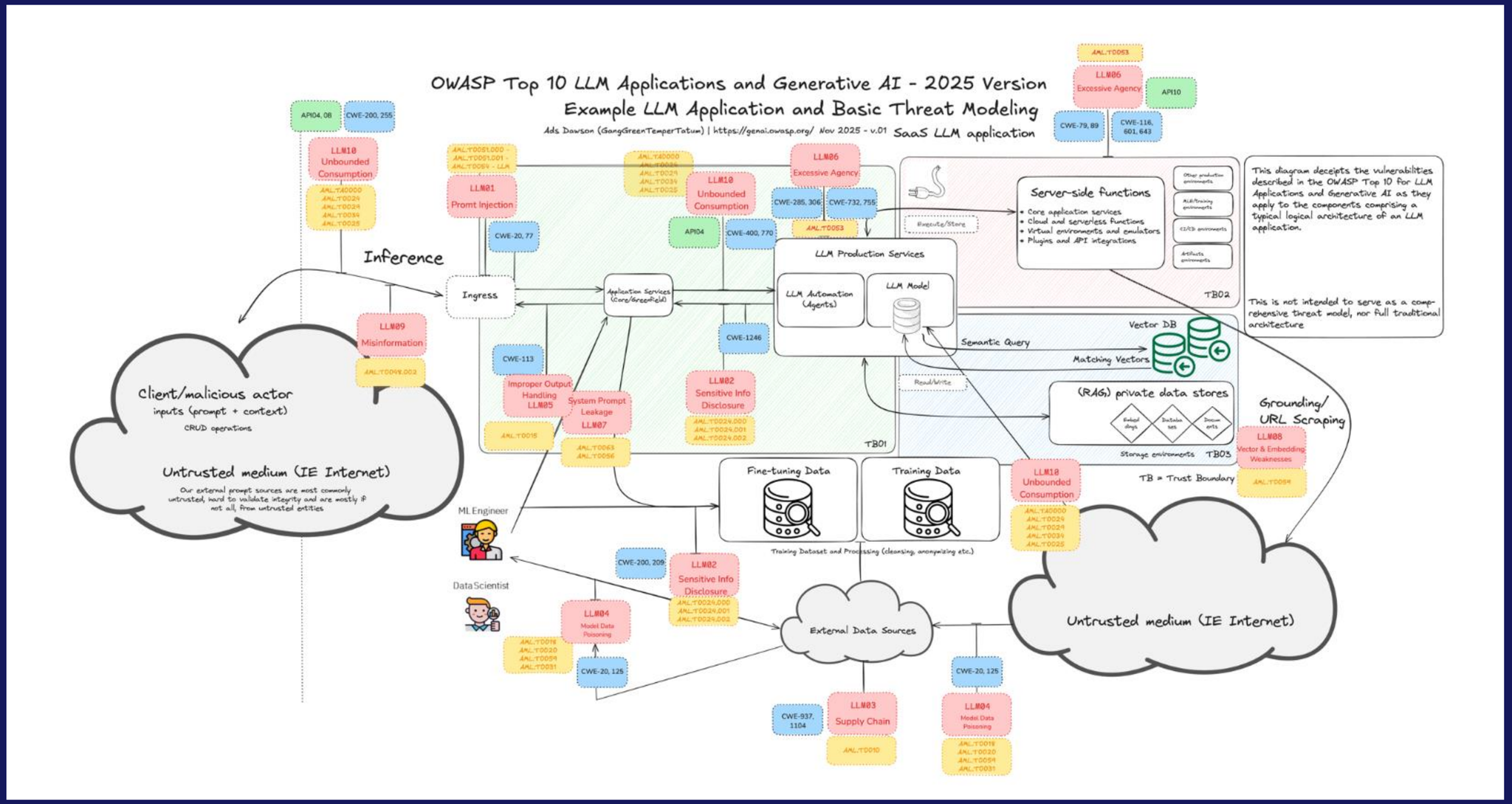
Naše Řešení a Mitigace:

- **Testování chování modelu** a jeho náchylnosti k dezinformacím.
- **Penetrační testy** simulující útoky na odepření služby.



ARICOMA

Architektura pro testování zranitelností LLM



Platforma pro vytváření inteligentních AI asistentů

- **Hardwarová vrstva:**
  - **Hosting v ČR:**
    - 24x GPU A100 80GB HGX
    - 3x Compute Node, 4x Storage Node, 1x Management node
  - **OnPremise v prostředí zákazníka:**
    - Servery s 1-8x GPU NVIDIA RTX Pro 6000 a vyšší
    - Výkonné CPU 16-128 jader a RAM 64GB – 2TB
    - Rychlá úložná kapacita: NVMe, SSD disky
- **Operační systém:**
  - Linux, Windows
- **Kontejnerizační platforma:**
  - Kubernetes, RedHat OpenShift

AI/ML infrastruktura:

- NVIDIA NIM + Nvidia AI Enteprise

- **Databázová vrstva:**
  - PostgreSQL + PGVector: Správa a vyhledávání vektorových dat
- **Orchestrace:**
  - MLFlow, LangFlow, n8n.oi

ARICOMA

Závěrečné Shrnutí: Jak bezpečně inovovat s AI?

- **AI = Dvojsečný meč:** Nové hrozby, nové příležitosti.
- **Aricoma = Inteligentní Obrana:** Proti AI útokům s AI obranou.
- **OWASP Top 10 = Náš Standard:** Bezpečnost od základu.
- **AI Act Ready:** Soulad s legislativou

Hostovaný, nebo On-Premise GPU Cluster

Platforma pro Multiagentní MLLM Systémy

ARICOMA

Děkujeme za pozornost

ARICOMA

aricoma.com

Hynek Cihlář
Aricoma System a.s.
hynek.cihlar@aricoma.com / +420 724 232 751