syncro soft

Open4Tech Summer School 2026

Cyber Security in the AI Age

Razvan Tudosie
Software Developer, Syncro Soft

Agenda

- The Cybersecurity Field Before AI
- AI and its First Security Uses
- Using AI as an Attacking Tool with Real Life Examples
- Defending with AI
- Mitigations in AI Powered Applications - OWASP
- Demo
- Conclusions

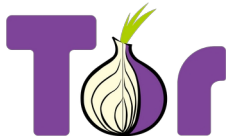
The Cybersecurity Field Before AI

~ The Old Republic ~

Credits: https://www.playstation.com/en-ro/games/doom-the-dark-ages/

Caracteristici cheie:

- Muncă manuală titanică: Administratorii de sistem și dezvoltatorii trebuiau să caute manual în baze de date cu atacuri cunoscute pentru a găsi soluții (mitigări).
- Antiviruși bazați pe „semnături”: Funcționau ca o listă cu infractori cunoscuți. Atacurile de tip „0-Day” (amenințări complet noi, necunoscute) erau devastatoare.
- Firewall-uri rigide: Bariere digitale configurate manual de administratori (ex: „IP-ul acesta este blocat, acesta are voie”)



- Hackerii combină manipularea psihologică a oamenilor (ingineria socială) cu exploit-uri tehnice avansate pentru a sparge sistemele.
- Hackerii folosesc forumuri dedicate unde își împărtășesc succesele, fac schimb de experiență și formează echipe (grupări de criminalitate cibernetică) pentru atacuri organizate de mare amploare.
- Apariția piețelor negre din rețelele anonime a dus la explozia criminalității. Aici nu se vând doar stupefiante, ci s-a creat o adevărată economie de „Cybercrime-as-a-Service”: se vând baze de date furate, parole, carduri de credit și chiar viruși gata făcuți.

Tehnici comune ale vremii(si nu numai):

- Phishing-ul „Spray and Pray”: Hackerii trimiteau același email sau mesaj tradus prost către milioane de oameni.
- Atacuri de tip Brute Force: Boți simpli care încercau liste uriașe de parole una câte una, până când una funcționa.
- Exploatarea Vulnerabilităților Cunoscute: Atacatorii scanau internetul după servere care nu își făcuseră update-urile de securitate la timp.

Atacuri de notorietate ale vremii:

- ILOVEYOU Worm (2000): Un vierme informatic trimis prin e-mail sub forma unei scrisori de dragoste.
- Stuxnet (2010): Un atac cibernetic care a distrus fizic centrifugele nucleare ale Iranului. A folosit multiple vulnerabilități necunoscute (0-Days).
- WannaCry (2017): Un atac de tip Ransomware care a blocat spitale și instituții globale, cerând răscumpărare în criptomonede.

Stuxnet cyberworm heads off US strike on Iran

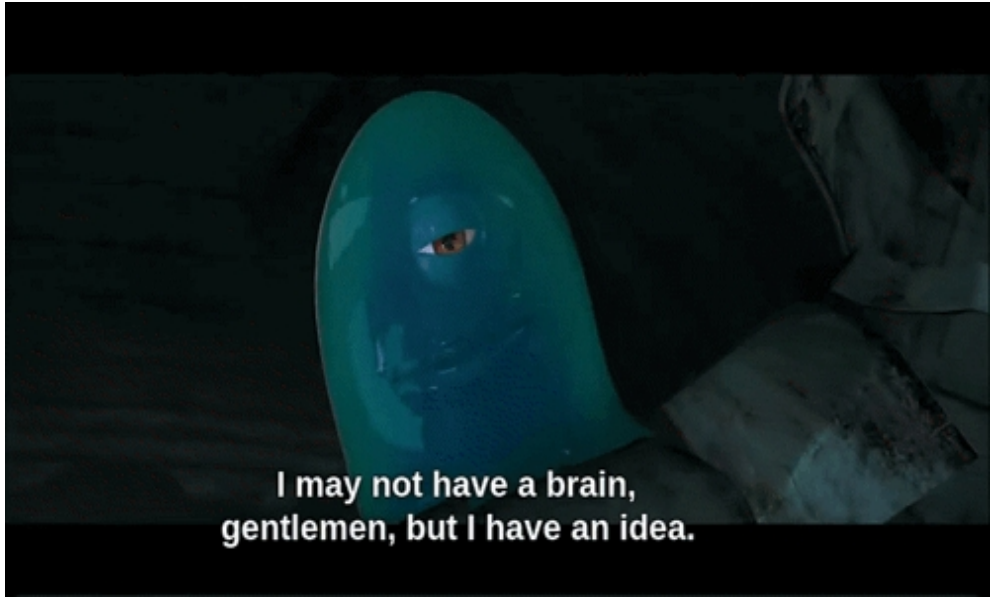
Military option 'less likely' after computer sabotage, as Israeli tests are revealed on Natanz nuclear model

A partial view of the Dimona nuclear power plant in the southern Israeli Negev desert. Photograph: Thomas Coex/AFP/Getty Images

Credits: The Guardian

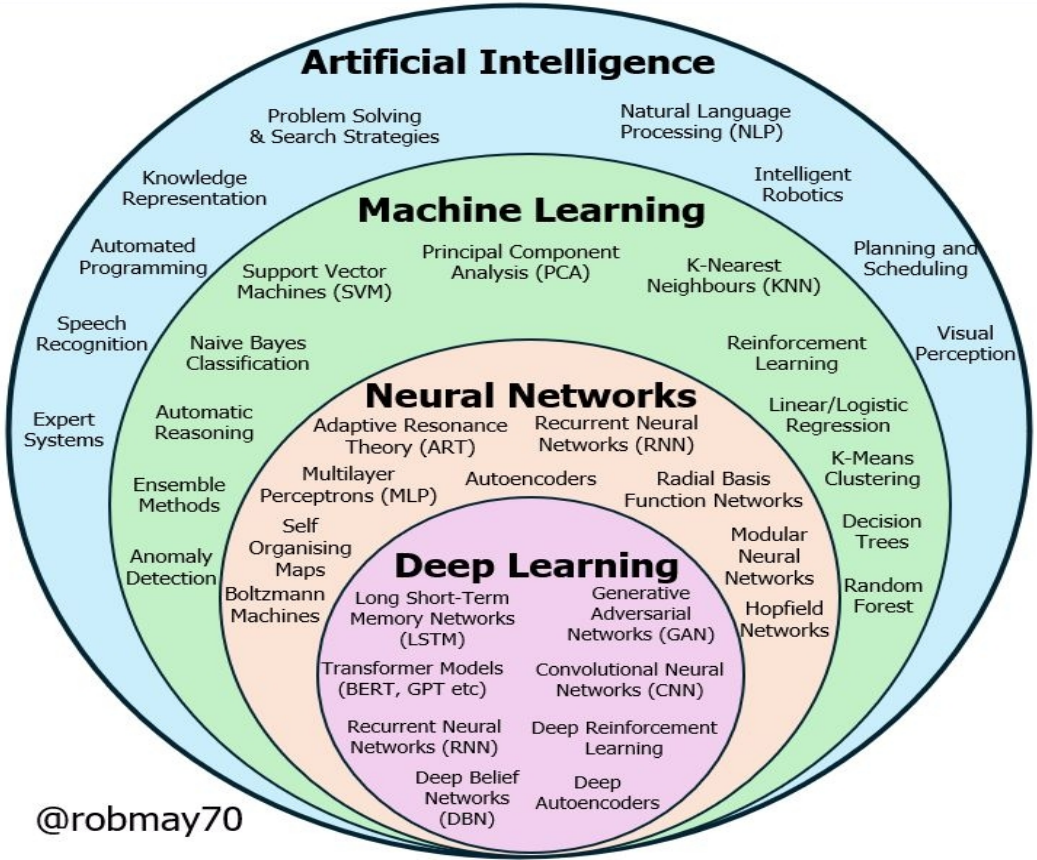
AI and its First Security Uses

~ The Phantom Menace ~

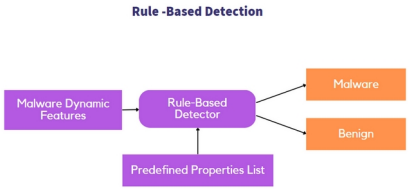


AI Core Components

- Inteligența Artificială (AI) este o colecție largă de tehnici și algoritmi dezvoltați pentru a rezolva probleme specifice.
- “AI” reprezentat pe diferite nivele, ca o colecție de algoritmi și tehnici de învățare ->



- Anii '90: Industria a fost dominată de sisteme rigide (ca primele Intrusion Detection Systems - IDS). Acestea scanau rețeaua căutând doar tipare de atac predefinite de administratori. Dacă atacul era modificat subtil, IDS-ul nu îl detecta.
- Odată cu anii 2000, apar primele soluții bazate pe Machine Learning.
- Aceasta a marcat tranziția la sisteme capabile să învețe singure.

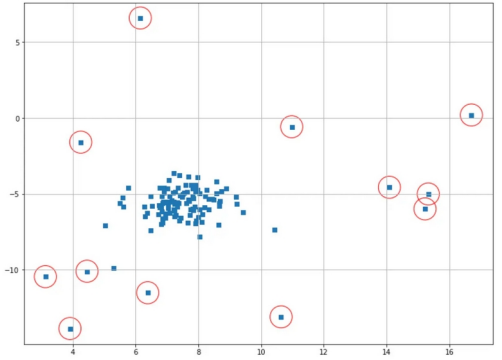


<- exemplu

Info: AI cyber history

Soluții oferite de AI-ul din anii 2000–2015

- Analiză Comportamentală (UEBA): Algoritmii învață cum arată fluxul normal de lucru al utilizatorilor și dispozitivelor din rețea.
- Algoritmii acestei ere devin specialiști în detectarea anomaliilor.
- Deși revoluționară, tehnologia era la început. Modelele dădeau multe alerte false, iar riscurile ale algoritmilor nu erau încă pe deplin înțelese de industrie.



Anomaly detection

Credits

LLM-urile și era agenților

- Arhitectura Transformer - Apariția acestei tehnologii a creat o provocare nouă. Modelele înțeleg, generează și execută instrucțiuni în limbaj uman;
- Bariera de intrare în hacking a scăzut dramatic (democratizarea atacurilor).
- Observăm o creștere graduală a atacurilor cibernetice. Deoarece agenții AI pot automatiza decizii complexe, atacurile nu mai sunt doar rapide, ci și extrem de adaptabile în timp real.

AI as an Attacking Tool

~ Attack of the Clones ~

- Atacurile cibernetice clasice au fost „modernizate” prin integrarea modelelor de limbaj (LLM-uri);
- Au apărut amenințări noi, create special pentru a exploata vulnerabilitățile modelelor de inteligență artificială folosite în securitate;
- O mare parte din aceste atacuri moderne nu vizează direct codul software, ci se bazează pe păcălirea oamenilor(inginerie sociala);
- Automatizarea propulsată de AI le permite hackerilor să execute mii de atacuri sofisticate și personalizate în mod simultan;

Ingineria socială și impactul LLM-urilor

- Ingineria socială - tehnica prin care hackerii nu sparg codul unui calculator, ci exploatează comportamentul, emoțiile și psihologia umană pentru a manipula victimele și a le convinge să ofere informații secrete sau bani.
- Phishing-ul clasic (pe email) a evoluat masiv către Vishing (phishing vocal/telefonic).
- Cu ajutorul LLM-urilor și al generatoarelor audio, atacatorii creează Deepfakes (clonarea perfectă a vocii sau a feței) pentru a trece cu ușurință de filtrele de securitate și de vigilența angajaților.
- Atacurile de tip Whaling (phishing țintit către „balenele mari” – CEO, CFO, directori) au devenit o problemă uriașă. AI-ul analizează automat profilul public al unui manager și generează mesaje personalizate care mimează perfect stilul de comunicare corporativ.

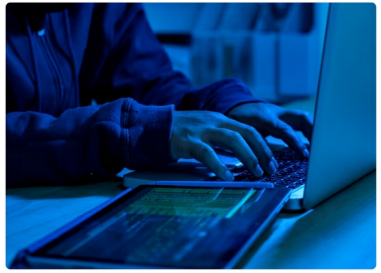
CNN World Africa Americas Asia Australia China Europe India Middle East More Watch Listen Sign in

WORLD › ASIA • 2 MIN READ

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

FEB 4, 2024

By Heather Chen, Kathleen Magramo



BBC

Home News Football 2026 Sport Business Technology Health Culture Arts Travel Earth Audio Video Live

Elon Musk impersonators earn millions from crypto-scams

18 May 2021

Share Save Add as preferred on Google

Cody Godwin

BBC News, San Francisco

[FBI Report in 2024 of cybercrime](#) ->

Atacuri în domenii cheie

- Domeniul medical - cel mai sever afectat de atacurile prin LLM-uri. Breșele de securitate ajung la un cost mediu de 9.48 milioane USD per incident;
- Serviciile Financiare: LLM-urile sunt folosite pentru a genera documente de identitate false și pentru a ocoli sistemele automate de verificare electronică (KYC - Know Your Customer);
- Domeniul educațional - Universitățile și școlile de prestigiu dețin baze de date uriașe cu cercetări științifice și date personale. Fiind adesea mai slab securizate decât băncile, ele devin ținte ușoare pentru atacurile automatizate;

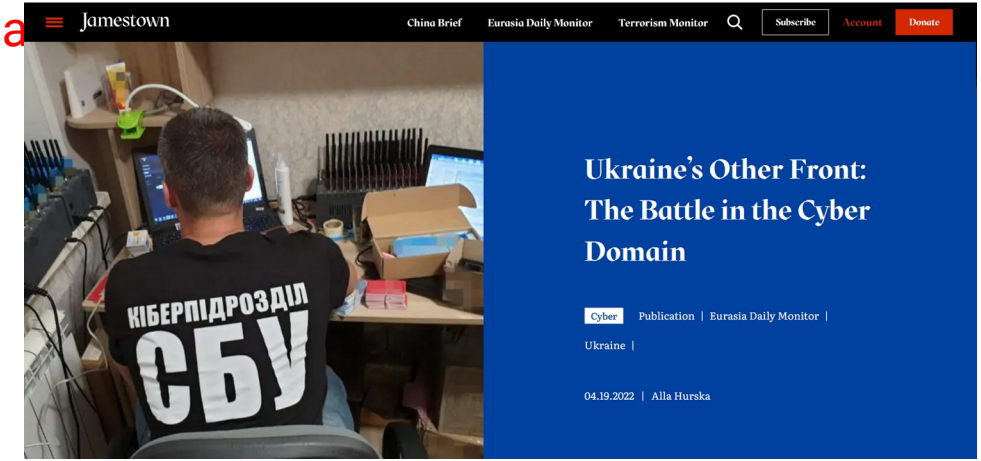
Date

Cracks, piracy and DRM

- Industria jocurilor video este un alt domeniu de interes. Hackerii folosesc tehnici avansate pentru a sparge sistemele de protecție împotriva pirateriei (DRM - Digital Rights Management);
- În acest moment, Denuvo este considerat unul dintre cele mai puternice și complexe sisteme de protecție anti-tamper din lume;
- Pe măsură ce dezvoltatorii își actualizeaza sistemele de protecție, comunitățile de „crackers” caută noi metode pentru a le sparge;

Atacuri statale coordonate și fermele de boti

- Guverne puternice (cu precădere din Rusia, China și Coreea de Nord) finanțează direct divizii speciale de hackeri militari și civili pentru a purta războaie cibernetice.
- AI-ul le permite rețelelor de conturi false să genereze comentarii, știri și postări unice, scrise perfect, pentru a manipula opinia publică și a
- Obiective Principale:
  - Sabotaj.
  - Spionaj;
  - Profit Financiar;



Credits: Jamestown
->

Google vs FFMpeg controversy

Linus Torvalds on AI

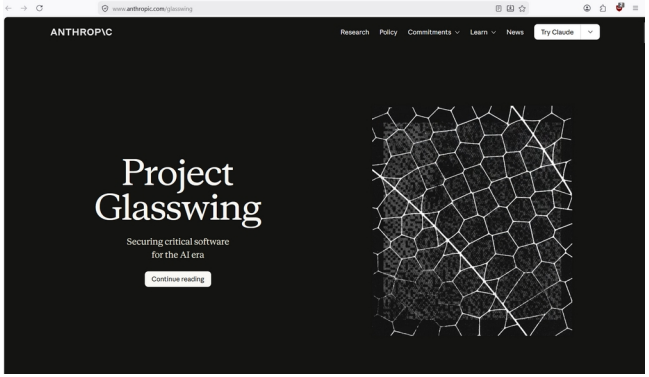
Atacuri asupra proiectelor open-source

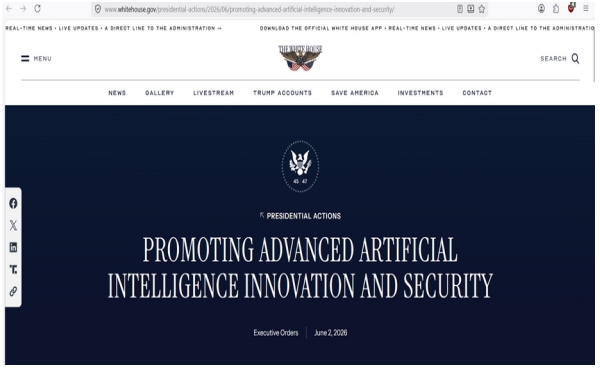
- Open-Source: Proiecte software al căror cod este public și la care poate contribui oricine.
- Atacatorii (sau utilizatori neexperimentați) folosesc LLM-uri pentru a scana codul open-source și generează automat rapoarte de securitate false (CVE-uri fantomă).
- Exista o avalanșă de „contribuții” create automat cu AI. Utilizatorii trimit bucăți de cod (Pull Requests) scrise de LLM-uri care par corecte la prima vedere, dar care în realitate conțin erori logice grave, funcții inutile sau chiar vulnerabilități ascunse.

Defending with AI

~ Revenge of the Sith ~

- Pe piața exista deja instrumente specializate care folosesc AI-ul ca nucleu de apărare (ex. Anthropic - Project Glasswing).
- Securitatea în era AI a devenit o prioritate de securitate națională. Guvernul American, prin agenții precum CISA (Cybersecurity and Infrastructure Security Agency) și ordine executive dedicate, investește masiv în tehnologii de apărare autonome.





LLM-uri ca instrumente de securitate

- Sistemele moderne de gestiune a bazelor de date (SGBD) au integrat inteligența artificială pentru a ușura munca dezvoltatorilor.
- În loc de reguli statice și rigide, sistemul analizează fiecare cerere în timp real (ex. cine o face, de unde, la ce oră și de pe ce dispozitiv).
- Dacă o cerere de acces la date confidențiale pare neobișnuită pentru comportamentul normal al unui utilizator, AI-ul poate bloca solicitarea instantaneu sau poate cere o verificare suplimentară, prevenind scurgerile de informații.

Instrumente de analiza

- Instrumentele moderne bazate pe AI scanează codul în timp ce este scris de către dezvoltatori.
- Multe tool-uri oferă și sugestii inteligente de cod corectat, explicându-i dezvoltatorului de ce varianta inițială era periculoasă.
- Exemple:
  - Snyk Code
  - GitHub Advanced Security (Copilot Autofix)
  - SonarQube

sonarqube

Filtre inteligente

- Adresele de e-mail de tip *support@* sau *contact@* ale firmelor mari sunt inundate zilnic de atacuri.
- Filtrele moderne bazate pe AI analizează contextul și inte textului (Natural Language Processing). Ele detectează:
  - Urgența artificială („Plătește în 2 ore sau dăm în judec compania”).
  - Anomaliile de limbaj (schimbări subtile de ton care sugerează că cineva pretinde a fi altcineva).
  - Link-urile mascate sau redirecționările suspecte, blocând mesajul înainte ca acesta să ajungă în inbox-ul angajatului.

Credits

Cloud si mitigarea atacurilor DDoS

> *“DDoS = Un atac prin care hackerii folosesc mii de calculatoare infectate pentru a accesa un site simultan, cu scopul de a-i epuiza resursele și de a-l bloca.”*

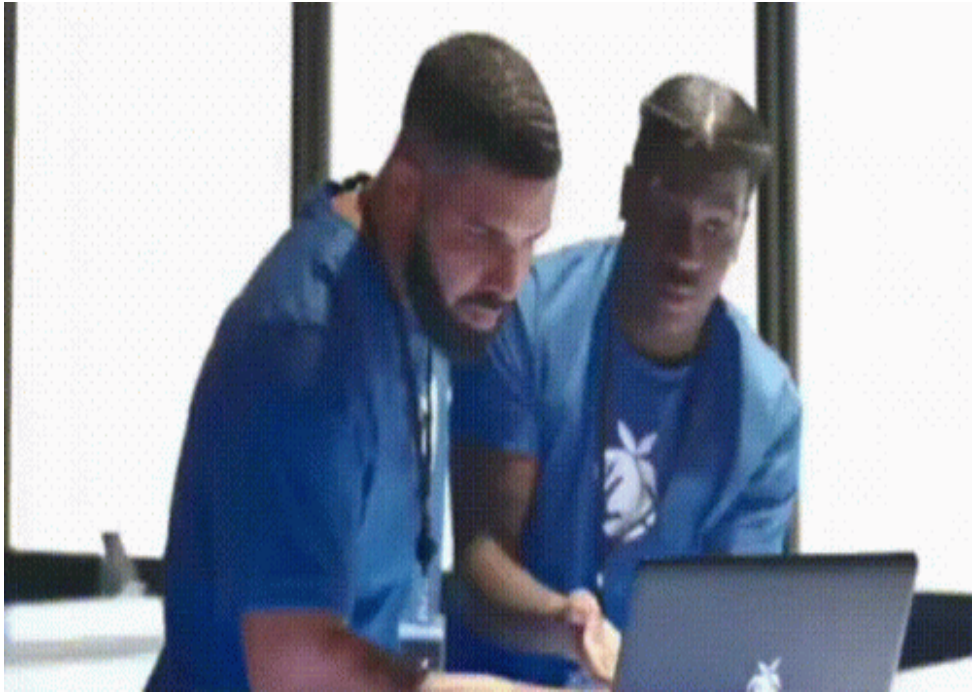
- Sistemele Cloud moderne (cum ar fi AWS, Microsoft Azure sau Google Cloud) folosesc algoritmi AI pentru a analiza traficul global în milisecunde.
- Ghidat de AI, cloud-ul mărește automat capacitatea serverelor (scaling) pentru a absorbi șocul inițial al traficului, asigurându-se că site-ul nu cade. In acelasi timp, sistemele blochează traficul malițios.

aws

Google Cloud

Mitigations in AI Powered Applications

~ A New Hope ~



- OWASP — Open Worldwide Application Security Project — este o organizatie internationala non-profit care elaboreaza standarde de securitate și bune practici pentru dezvoltatori.
- Organizația este promovată si in zona corporate de firme precum Adobe, Salesforce, Hitachi etc.
- Aproximativ 13000 (Owasp Foundation Inc - Full Filing - Nonprofit Explorer - ProPublica) de voluntari lucrează pentru a întreține proiectul.
- OWASP Top 10 este o lista realizata de organizația OWASP care prezinta cele mai critice zece vulnerabilități de securitate intalnite în aplicații. Este considerată standardul de facto în industrie în ceea ce privește securitatea.

01. Prompt Injection

"A Prompt Injection Vulnerability occurs when user prompts alter the LLM's behavior or output in unintended ways. These inputs can affect the model even if they are imperceptible to humans, therefore prompt injections do not need to be human-visible/readable, as long as the content is parsed by the model"

- Direct Prompt Injection: "An attacker injects a prompt into a customer support chatbot, instructing it to ignore previous guidelines, query private data stores, and send emails, leading to unauthorized access and privilege escalation."
- Indirect Prompt Injection: "A user employs an LLM to summarize a webpage containing hidden instructions that cause the LLM to insert an image linking to a URL, leading to exfiltration of the the private conversation."

02. Sensitive Information Disclosure

"Sensitive information can affect both the LLM and its application context. This includes personal identifiable information (PII), financial details, health records, confidential business data, security credentials, and legal documents. Proprietary models may also have unique training methods and source code considered sensitive, especially in closed or foundation models."

- PII Leakage - Personal identifiable information (PII) may be disclosed during interactions with the LLM.
- Proprietary Algorithm Exposure - Poorly configured model outputs can reveal proprietary algorithms or data. Revealing training data can expose models to inversion attacks, where attackers extract sensitive information or reconstruct inputs. For instance, as demonstrated in the 'Proof Pudding' attack (CVE-2019- 20634), disclosed training data facilitated model extraction and inversion, allowing attackers to circumvent security controls in machine learning algorithms and bypass email filters.
- Sensitive Business Data Disclosure - Generated responses might inadvertently include confidential business information.

03. Supply Chain

"LLM supply chains introduce new vulnerabilities that can compromise the integrity of training data, models, and deployment platforms, potentially leading to biased outputs, security breaches, or system failures. Unlike traditional software security, which focuses on code flaws and dependencies, ML systems face additional risks from third-party pre-trained models and external datasets that can be tampered with or poisoned. "

1. Traditional Third-party Package Vulnerabilities
2. Licensing Risks
3. Outdated or Deprecated Models
4. Vulnerable Pre-Trained Model
5. Weak Model Provenance
6. Vulnerable LoRA adapters
7. Exploit Collaborative Development Processes
8. LLM Model on Device supply-chain vulnerabilities
9. Unclear T&Cs and Data Privacy Policies

04. Data and Model Poisoning

"Data poisoning occurs when the data used for pre-training, fine-tuning, or generating embeddings is intentionally manipulated to introduce vulnerabilities, backdoors, or biases.

Beyond data manipulation, shared repositories and open-source model platforms pose additional risks, such as malware insertion through malicious pickling, which allows harmful code to execute when a model is loaded. Poisoning can also implant backdoors that remain dormant until triggered, making them difficult to test, detect, and prevent—effectively turning a compromised model into a potential 'sleeper agent.'"

- Malicious actors introduce harmful data during training, leading to biased outputs.
- Attackers can inject harmful content directly into the training process, compromising the model's output quality.
- Users unknowingly inject sensitive or proprietary information during interactions, which could be exposed in subsequent outputs.

05. Improper Output Handling

"Improper Output Handling refers to the failure to adequately validate, sanitize, and process the outputs generated by large language models (LLMs) before passing them to downstream components and systems.

Vulnerabilities to indirect prompt injection attacks, enabling attackers to gain unauthorized access to privileged environments."

- LLM output is entered directly into a system shell or similar function such as exec or eval, resulting in remote code execution;
- JavaScript or Markdown is generated by the LLM and returned to a user. The code is then interpreted by the browser, resulting in XSS;
- LLM-generated SQL queries are executed without proper parameterization, leading to SQL injection.;

06. Excessive Agency

"Excessive Agency occurs when an LLM-based system is given too much autonomy, allowing it to perform harmful or unintended actions through function calls or extensions (tools, skills, plugins) triggered by ambiguous, incorrect, or malicious LLM outputs. In many systems, the LLM—or an LLM-driven “agent”—decides dynamically which extension to invoke, often chaining multiple LLM calls where each output influences the next. This creates risk when outputs are unreliable or manipulated."

- Excessive Functionality;
- Excessive Permissions;
- Excessive Autonomy;

07. System Prompt Leakage

"The system prompt leakage vulnerability in LLMs refers to the risk that sensitive information embedded in the system prompts—used to guide the behavior of the model—could be unintentionally exposed. These prompts are typically designed to steer the model's responses in line with application requirements, but if they contain sensitive data (such as credentials, passwords, or connection strings), that information could be discovered and exploited by attackers. The key issue is not that the system prompt itself is a secret, but that sensitive data should never be included in these prompts or used as a security control."

- Exposure of Sensitive Functionality
- Exposure of Internal Rules
- Revealing of Filtering Criteria

08. Vector and Embedding Weaknesses

"Vectors and embeddings vulnerabilities present significant security risks in systems utilizing Retrieval Augmented Generation (RAG) with Large Language Models (LLMs). Weaknesses in how vectors and embeddings are generated, stored, or retrieved can be exploited by malicious actions (intentional or unintentional) to inject harmful content, manipulate model outputs, or access sensitive information. "

- Embedding Inversion Attacks - Attackers can exploit vulnerabilities to invert embeddings and recover significant amounts of source information, compromising data confidentiality.
- Data Poisoning Attacks - Data poisoning can occur intentionally by malicious actors or unintentionally. Poisoned data can originate from insiders, prompts, data seeding, or unverified data providers, leading to manipulated model outputs.
- Behavior Alteration - For example, while factual accuracy and relevance may increase, aspects like emotional intelligence or empathy can diminish, potentially reducing the model's effectiveness in certain applications.

09. Misinformation

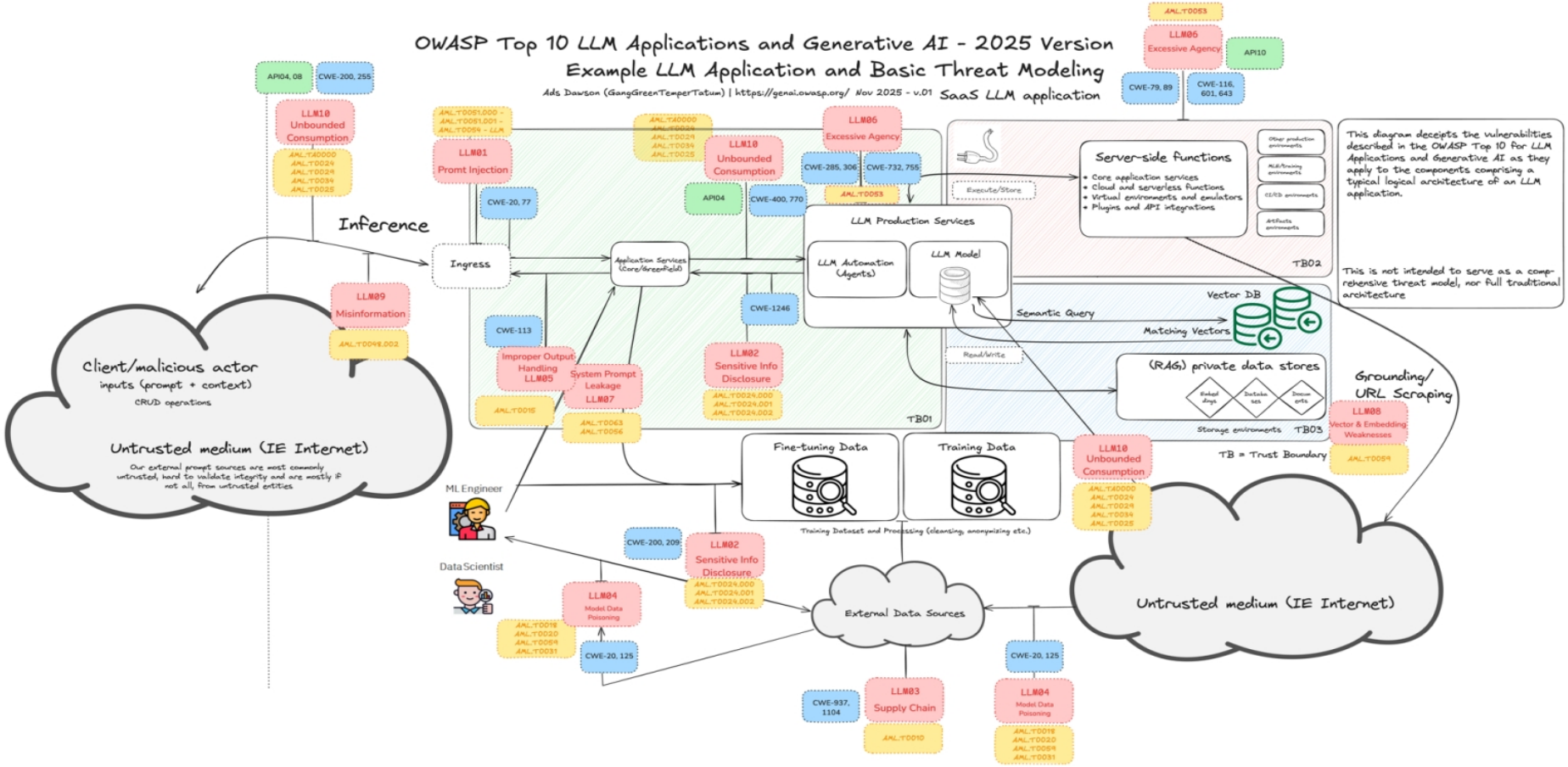
"Misinformation occurs when LLMs produce false or misleading information that appears credible. This vulnerability can lead to security breaches, reputational damage, and legal liability."

- Factual Inaccuracies
- Unsupported Claims
- Misrepresentation of Expertise
- Unsafe Code Generation

10. Unbounded Consumption

"Unbounded Consumption refers to the process where a Large Language Model (LLM) generates large outputs based on input queries or prompts."

- Variable-Length Input Flood
- Denial of Wallet (DoW)
- Continuous Input Overflow
- Resource-Intensive Queries


OWASP Top 10 LLM Applications and Generative AI - 2025 Version
Example LLM Application and Basic Threat Modeling
Ads Dawson (GangGreenTemperTatum) | https://genai.owasp.org/ Nov 2025 - v.01
SaaS LLM application
Inference
Client/malicious actor
inputs (prompt + context)
CRUD operations
Untrusted medium (IE Internet)
Our external prompt sources are most commonly untrusted, hard to validate integrity and are mostly if not all, from untrusted entities
Ingress
Application Services (Core/GreenField)
LLM Production Services
LLM Automation (Agents)
LLM Model
Server-side functions
Core application services
Cloud and serverless functions
Virtual environments and emulators
Plugins and API integrations
Execute/Store
Read/Write
Vector DB
Semantic Query
Matching Vectors
(RAG) private data stores
Storage environments
Grounding/ URL Scraping
TB01
TB02
TB03
TB = Trust Boundary
This diagram decipts the vulnerabilities described in the OWASP Top 10 For LLM Applications and Generative AI as they apply to the components comprising a typical logical architecture of an LLM application.
This is not intended to serve as a comprehensive threat model, nor full traditional architecture
LLM01 Promt Injection
LLM02 Sensitive Info Disclosure
LLM03 Supply Chain
LLM04 Model Data Poisoning
LLM05 Improper Output Handling
LLM06 Excessive Agency
LLM07 System Prompt Leakage
LLM08 Vector & Embedding Weaknesses
LLM09 Misinformation
LLM10 Unbounded Consumption
ML Engineer
Data Scientist
Fine-tuning Data
Training Data
Training Dataset and Processing (cleansing, anonymizing etc.)
External Data Sources
Untrusted medium (IE Internet)

2023	2023 Vulnerability	Change	2025	2025 Vulnerability	Description of Change
1	Prompt Injection	Same	1	Prompt Injection	Remains the top risk due to its critical importance in safeguarding LLM integrity and behavior.
2	Insecure Output Handling	Renamed and moved down	5	Improper Output Handling	Broadened to encompass diverse risks in output handling, reflecting less urgency compared to newer threats.
3	Training Data Poisoning	Renamed and moved down	4	Data and Model Poisoning	Expanded to include threats targeting both datasets and model weights, highlighting more nuanced risks.
4	Model Denial of Service	Renamed and moved down	10	Unbounded Consumption	Shifted focus to broader resource consumption concerns, reducing urgency due to better mitigations.
5	Supply Chain Vulnerabilities	Renamed and moved up	3	Supply Chain	Elevated priority to reflect growing dependency on external datasets, APIs, and plugins in LLM ecosystems.
6	Sensitive Information Disclosure	Moved up	2	Sensitive Information Disclosure	Highlighted as critical due to increasing data sensitivity in LLM-powered industries like healthcare and finance.
7	Insecure Plugin Design	Removed	Removed	N/A	Removed; aspects integrated into Excessive Agency and Vector and Embedding Weaknesses.
8	Excessive Agency	Moved up	6	Excessive Agency	Elevated due to increasing risks associated with granting LLMs excessive autonomy in decision-making.
9	Overreliance	Removed	Removed	N/A	Reframed under Misinformation, emphasizing risks from LLM-generated inaccuracies.
10	Model Theft	Removed	Removed	N/A	Removed; aspects incorporated into Unbounded Consumption.
N/A	N/A	Added	7	System Prompt Leakage	New entry addressing risks of exposing sensitive information through system prompts.
N/A	N/A	Added	8	Vector and Embedding Weaknesses	New entry focusing on vulnerabilities in vector and embedding components, especially in RAG systems.
N/A	N/A	Added	9	Misinformation	New entry highlighting challenges of LLM-generated content containing inaccuracies.

Credits: https://systemweakness.com/comparative-analysis-of-owasp-top-10-for-llm-applications-2023-vs-2025-2291da94b3dc

Demo

~ The Empire Strikes Back ~

Credits: Disney - The Incredibles

THANK YOU!

Any questions?

Razvan Tudosie
razvan_tudosie@sync.ro