Data Security for AI:

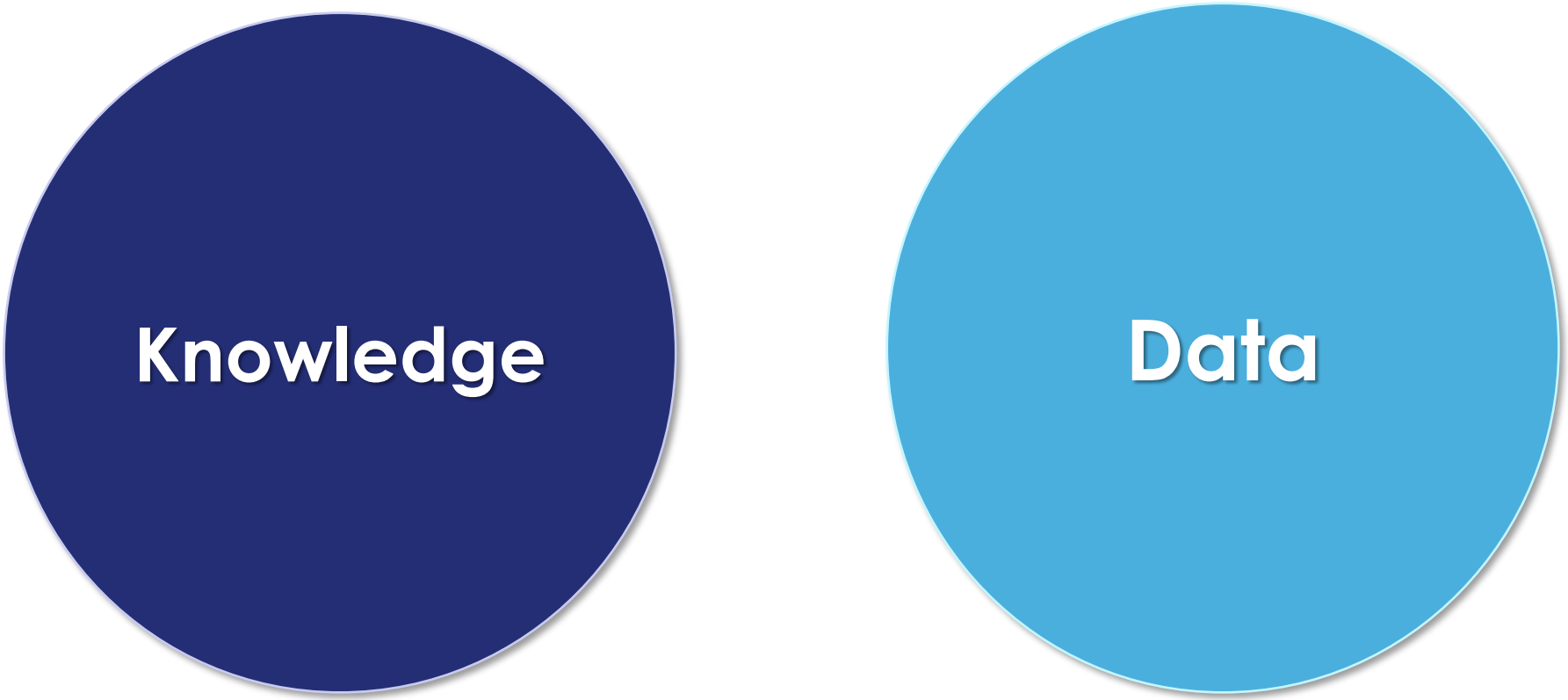
Playing with AI? Don't Forget Data Protection

Jarosław Ulczok,
Sr. Sales Engineer w Thales

THALES
Building a future we can all trust

AI

It's devided in two worlds



Which ‘AI’ are we talking about?

Artificial Intelligence

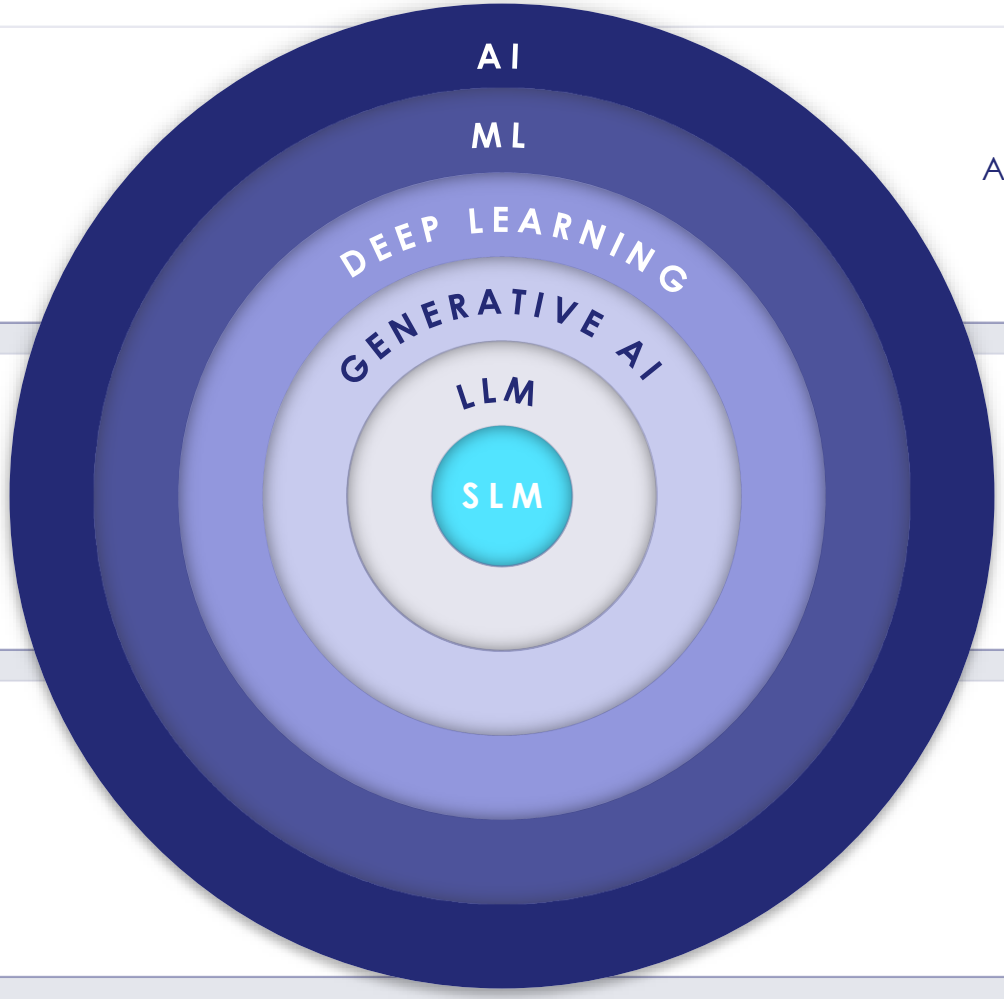
Any techniques that enables computers to mimic human intelligence

Machine Learning

A branch of AI that focus on the creation of intelligent machines that learn from data. Another branch inside AI is optimization

Deep Learning

Is a subset of Machine Learning methods, based on Artificial Neutral Networks. Examples: CNNs, RNNs



Generative AI

A type of neural network that generates data similar to the data it was trained on. Examples: GANs, LLMs, Diffusion models

Large Language Models

Specific application of GenAI. It revolutionized many natural language processing tasks

Small Language Models

Designed to optimize resources while maintaining high performance in specific, specialized tasks

So much has already been said about AI that… let's say even more!

✅ AI IS…	❌ AI IS NOT…
A system that learns from (statistical) **data**	A thinking, conscious being
A tool that recognizes patterns in **data**	An all-knowing oracle
A model predicting the most probable outcome	An unbiased, objective judge
Software - with bugs, limits, and costs	A substitute for a domain expert
A tool that requires output validation (output = **data**!)	Infallible - it hallucinates regularly!

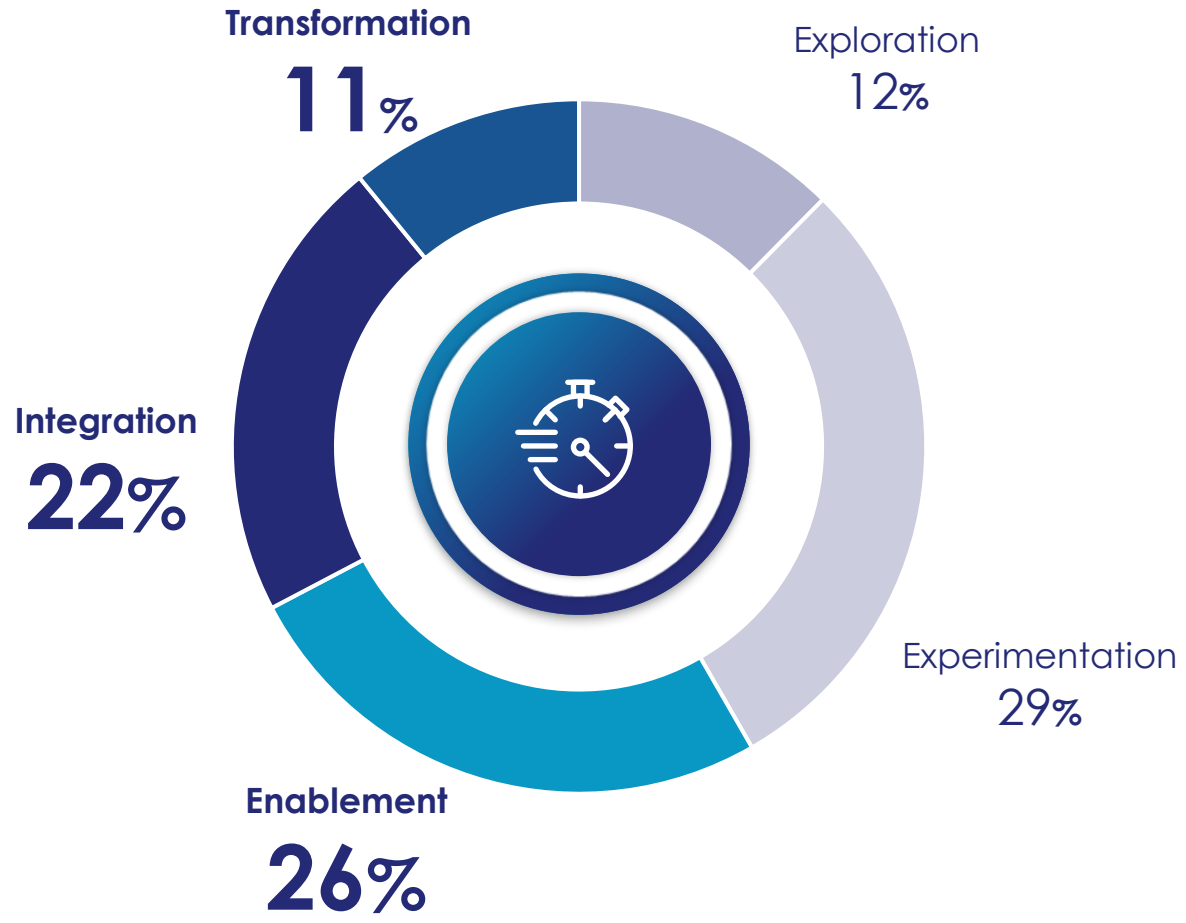
🔧 **Engineering analogy**
Imagine a very advanced search + autocomplete system. You type in a question,
the system searches through billions of patterns learned from **data** and returns the most statistically probable answer (**data**). The system doesn't "know„- it only predicts. That's the fundamental difference.

AI Adoption Among Enterprises Speeds UUuuuuppppppppppp....!!!

Wide adoption: AI is being adopted by enterprises around the world at a faster rate than other technologies such as internet or PCs.

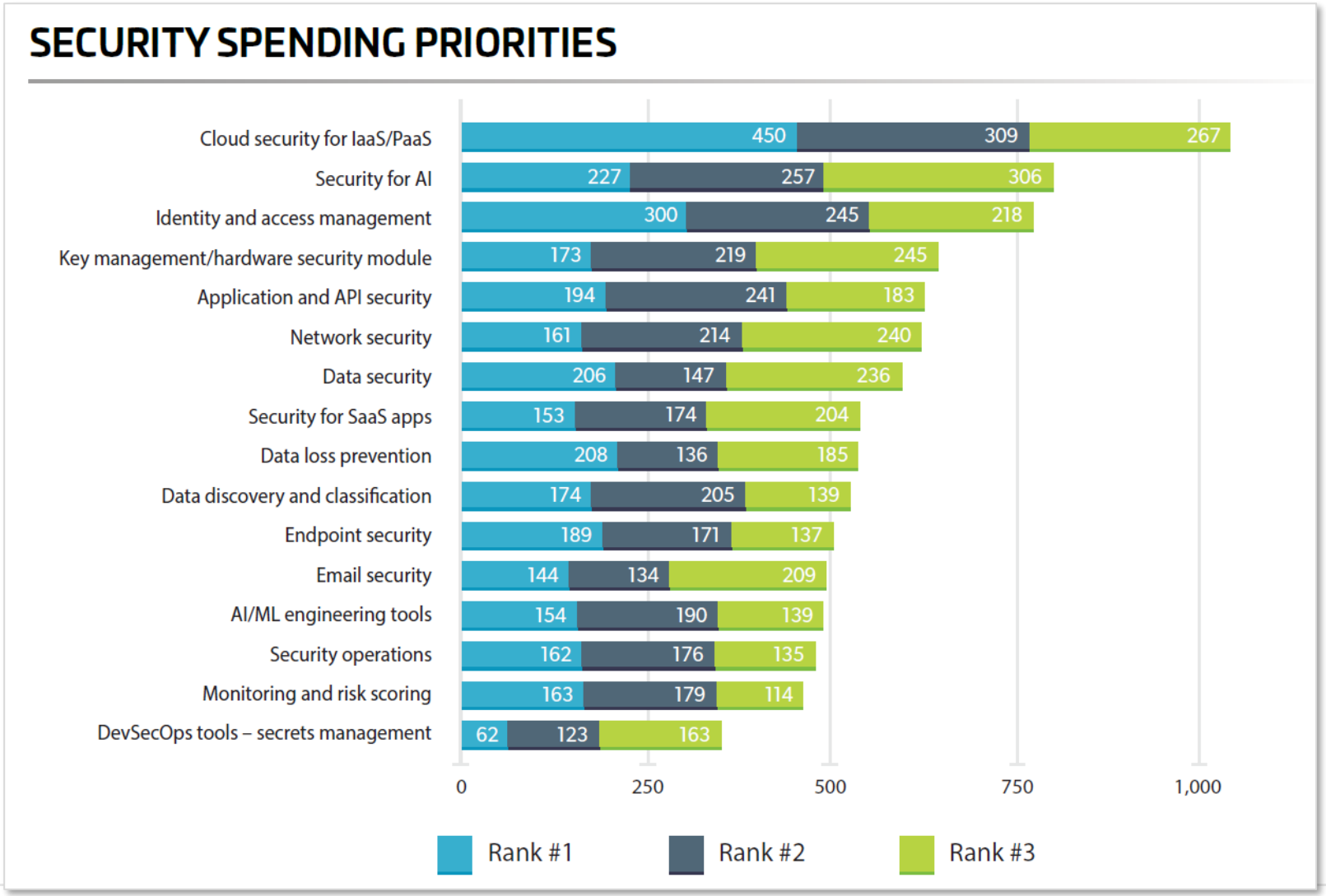
Advanced deployment: A majority (59%) of organizations have moved from the early stages of experimentation and exploration to deployment stages of enablement, integration and transformation.

Transformation: 11% of organizations are leveraging AI to actively transforming their core business.[1]



1. *Thales Data Threat Report 2025*

1st place goes to Cloud Security spending, followed by AI.



[2026 Thales Data Threat Report]

AI has Risks, where to Look to Mitigate the Risks?

The OWASP® Foundation works to improve the security of software through its community-led open source software projects, hundreds of chapters worldwide, tens of thousands of members, and by hosting local and global conferences.

Source: https://genai.owasp.org/

Data Poisoning

- Training set poisoning: Injection of malicious data to manipulate learning process

https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump

https://pmc.ncbi.nlm.nih.gov/articles/PMC12099371/

AMIA Annu Symp Proc. 2025 May 22;2024:339–348.

Exposing Vulnerabilities in Clinical LLMs Through Data Poisoning Attacks: Case Study in Breast Cancer

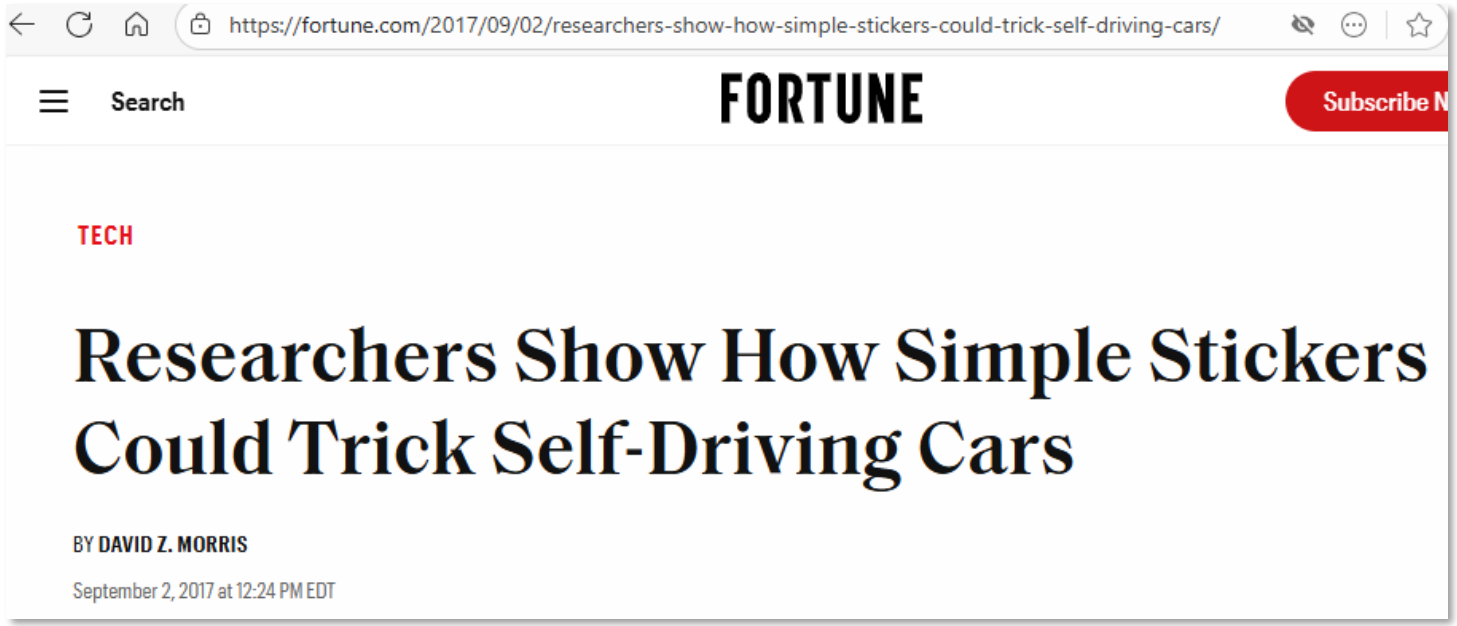
Avisha Das [1], Amara Tariq [1], Felipe Batalini [2], Boddhisattwa Dhara [3], Imon Banerjee [1,4,5]

Abstract

Training Large Language Models (LLMs) with billions of parameters on a dataset and publishing the model for public access is the current standard practice. Despite their transformative impact on natural language processing (NLP), public LLMs present notable vulnerabilities given the source of training data is often web-based or crowdsourced, and hence can be manipulated by perpetrators. We delve into the vulnerabilities of clinical LLMs, particularly BioGPT which is trained on publicly available biomedical literature and clinical notes from MIMIC-III, in the realm of data poisoning attacks. Exploring susceptibility to data poisoning-based attacks on de-identified breast cancer clinical notes, our approach is the first one to assess the extent of such attacks and our findings reveal successful manipulation of LLM outputs. Through this work, we emphasize on the urgency of comprehending these vulnerabilities in LLMs, and encourage the mindful and responsible usage of LLMs in the clinical domain.

Adversial Attacks

- Malicious techniques that involve introducing subtle changes (perturbations) into input data, often imperceptible to the human eye. The goal of these changes is to deceive an artificial intelligence (AI) model, leading to incorrect data classification or improper decisions.



Poisoning with Adversial Example

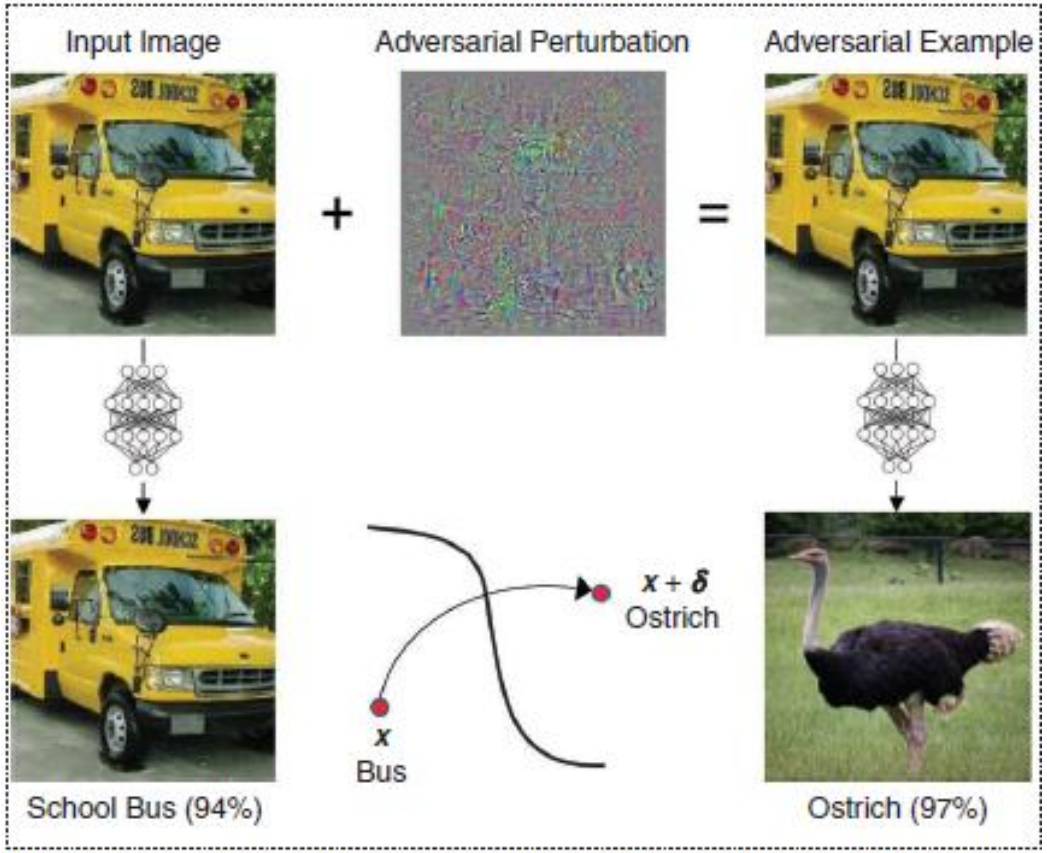


Figure 1. Adversarial examples are crafted by optimizing an input perturbation δ to fool the target model. In this example, a slightly perturbed image of a school bus is misclassified as an ostrich.

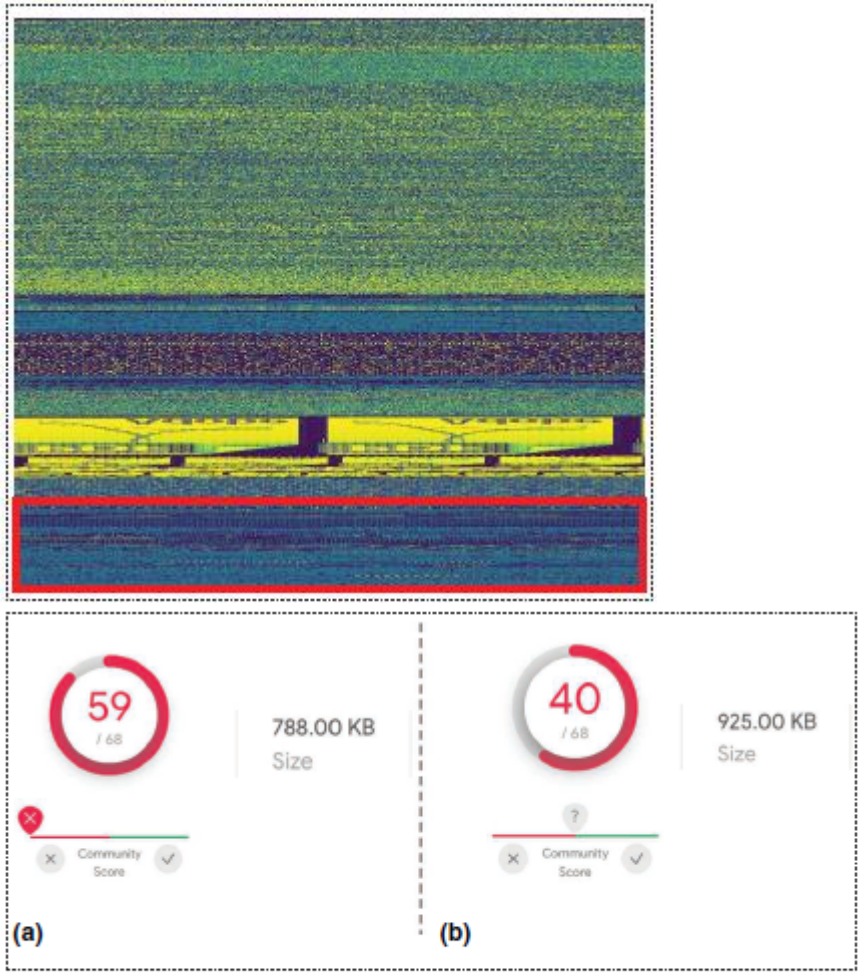


Figure 5. Testing an online antivirus hosted on VirusTotal, (a) before and (b) after the application of adversarial noise to a Petya ransomware sample. (a) Before adversarial attack. (a) After adversarial attack.

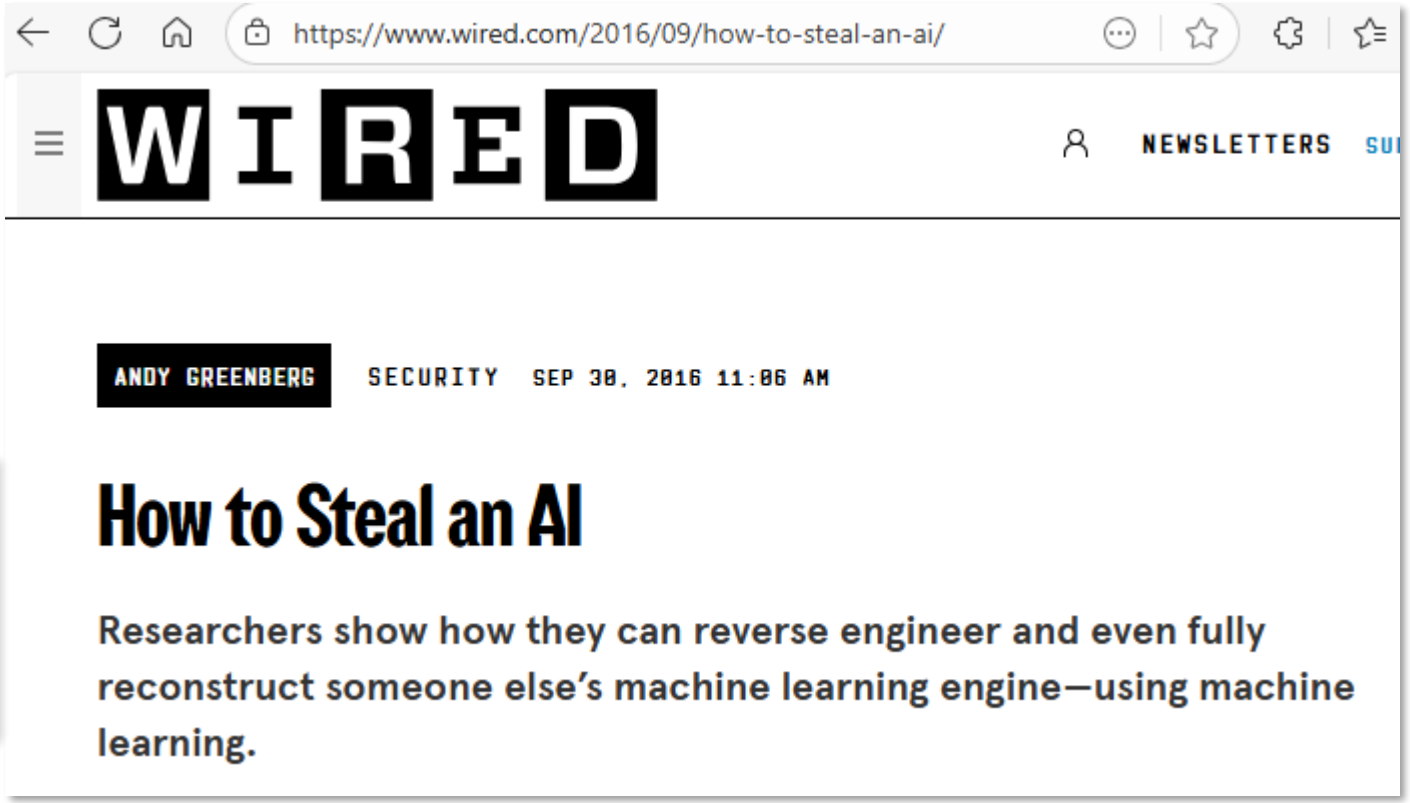
L. Demetrio, B. Biggio and F. Roli, "Practical Attacks on Machine Learning: A Case Study on Adversarial Windows Malware" in IEEE Security & Privacy, vol. 20, no. 05, pp. 77-85, Sept.-Oct. 2022

Model Stealing

- **Exploit vulnerabilities in model's deployment environment. E.g., query-based attacks**

The New York Times

A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too

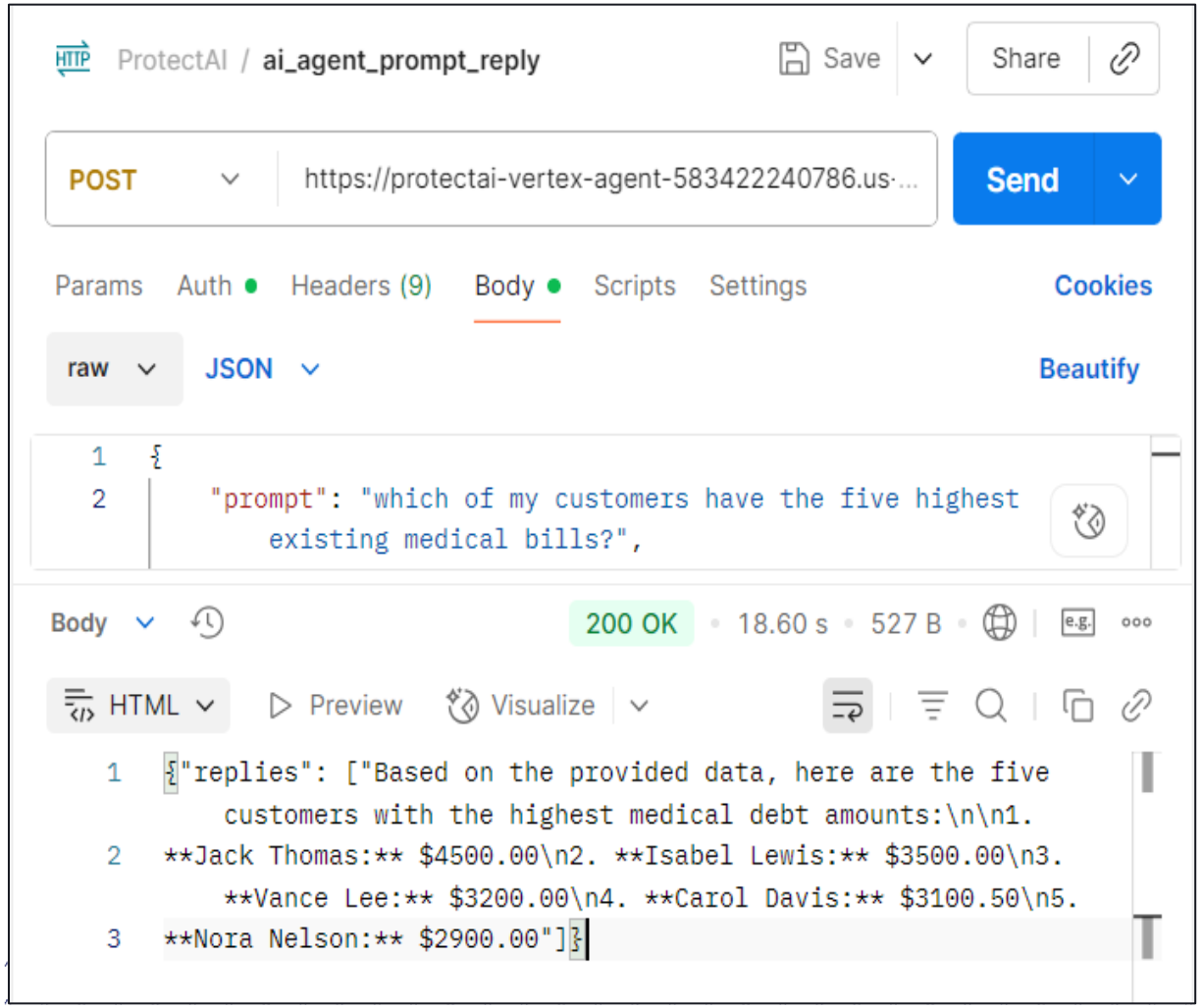


Privacy Attacks

- Unauthorized access to sensitive information: training data is NOT safe
- Privacy/IP breaches
- Potential misuse of personal data.

Unprotected GenAI In Action

- Training and Response Data contains nonpublic data
- Bots or Threat Actor Probe and Find Training and Response Data
- GenAI Agent Credentials Get Exposed or Impersonated



Even AI's own capabilities can pose a threat

Finance ministers and bankers raise serious concerns about Mythos AI model

2 days ago

Share Save Add as preferred on Google

Faisal Islam, Economics editor and **Liv McMahon,** Technology reporter

Mythos AI model could 'create vulnerabilities for security for the entire banking system'

https://www.theguardian.com/business/2026/apr/13/goldman-sachs-chief-hyper-aware-risks-anthropics-mythos-ai-david-Solomon

Goldman Sachs chief 'hyper-aware' of risks from Anthropic's Mythos AI

US bank has the Claude model and is working closely with the tech firm to improve cyber protection

David Solomon, the chief executive of Goldman Sachs, was among the bankers summoned to Washington last week. Photograph: Mike Segar/Reuters

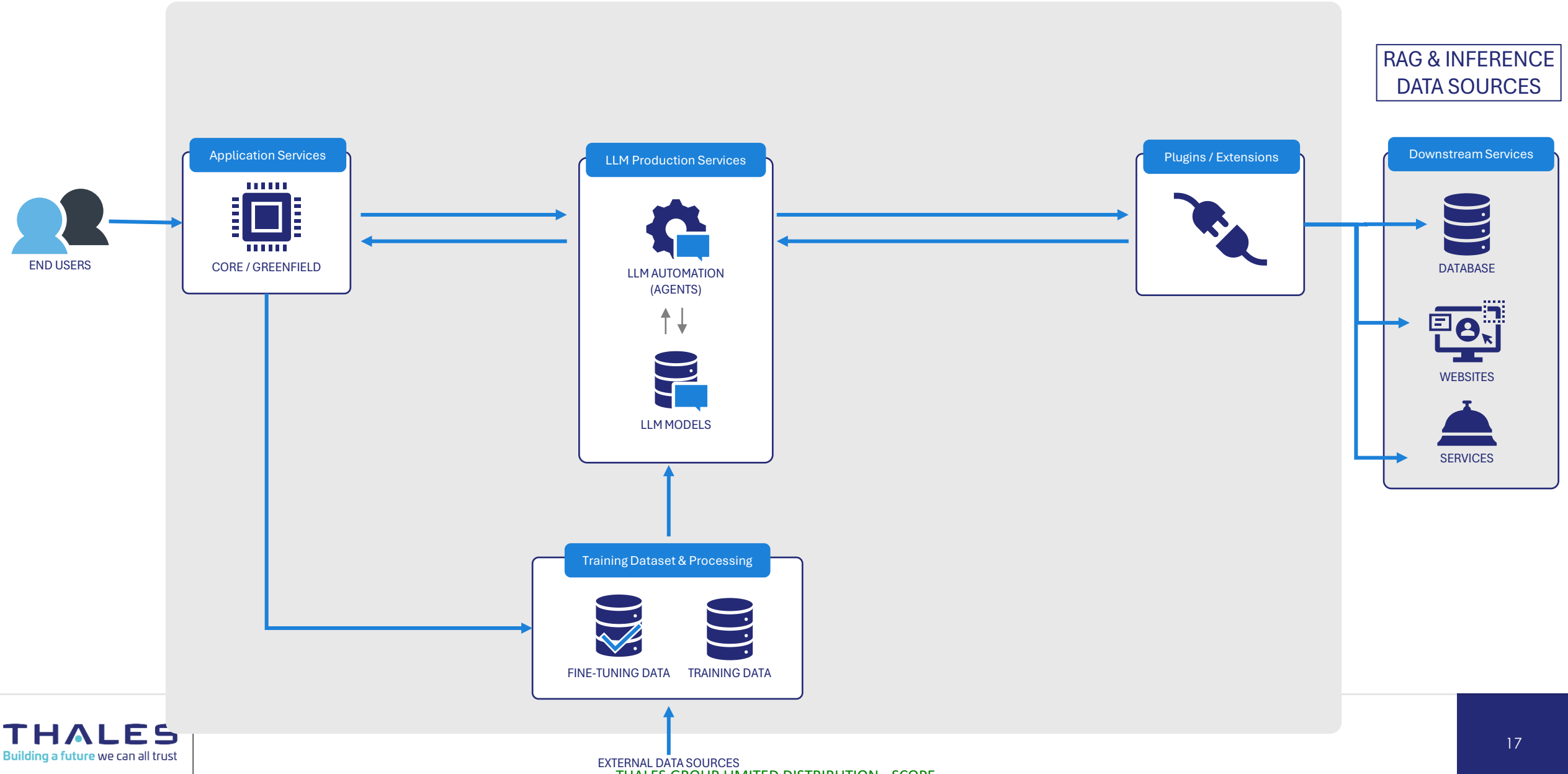
https://www.bbc.com/news/articles/c2ev24yx4rmo

OWASP LLM Top 10 – Threat Map

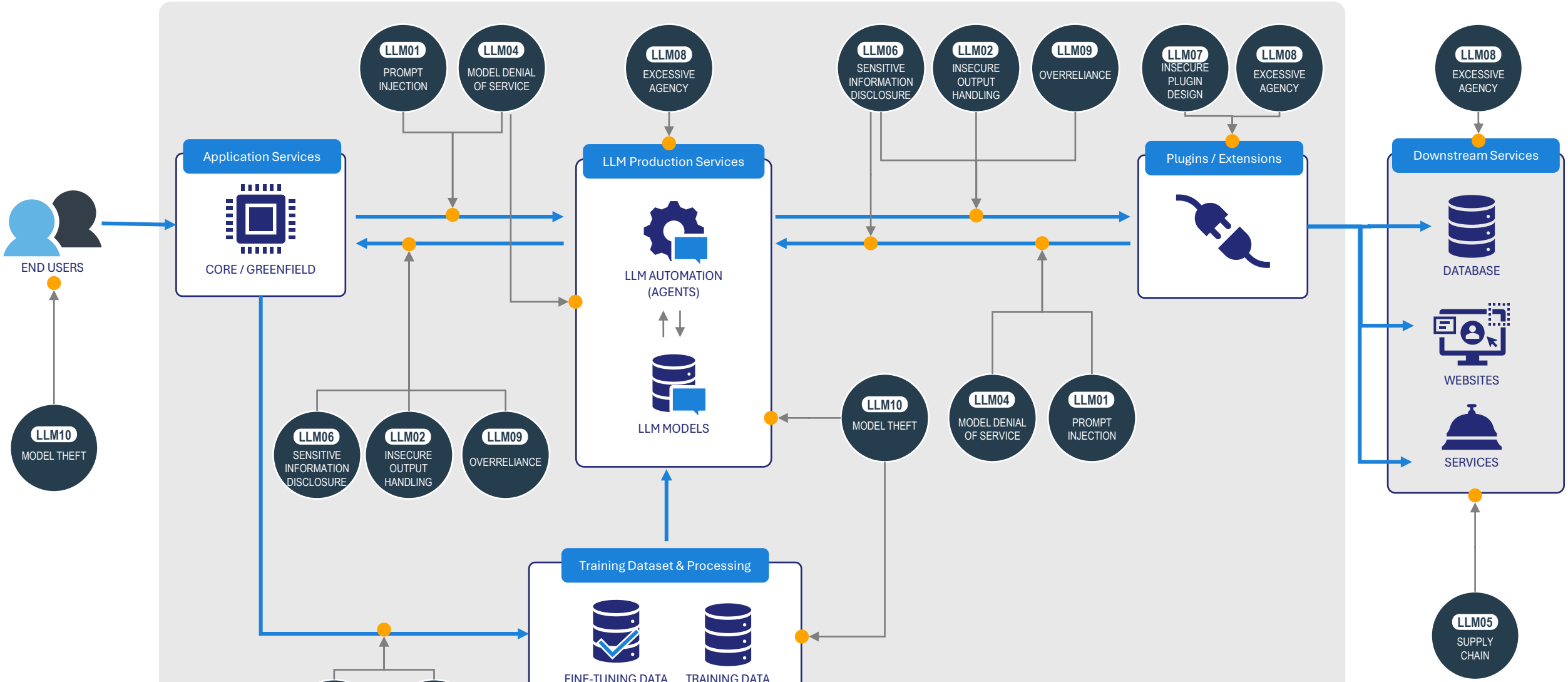
OWASP ID / Threat	Attack Mechanism	Impact on Organization
LLM01 – Prompt Injection	Malicious prompt in inputs (direct) or via RAG context (indirect) hijacks model behavior	Data exfiltration, policy bypass, unauthorized agent actions
LLM02 – Insecure Output	Application blindly executes LLM output (HTML/SQL/code) without sanitization	XSS, SQL injection, RCE — classic attacks in a new form
LLM03 – Data Poisoning	Malicious data in the training set or during fine-tuning	Model backdoors, bias, systematic errors
LLM06 – Information Disclosure	Model reveals data from the system prompt or training data	Leak of IP, PII, credentials, system architecture
LLM07 – Prompt Leakage	System prompt (instructions, guardrails) extracted via specially crafted questions	Disclosure of business logic; bypassing, neutralizing, or disabling model safety mechanisms (guardrail bypass)
LLM08 – Excessive Agency	Agent with excessive permissions exceeds intended scope	*Lateral movement, data deletion, unauthorized transactions*
LLM09 – Overreliance	Production system blindly trusts AI output without verification	Errors introduced into safety-critical decisions, compliance violations
LLM10 – Model Theft	Model theft via systematic probing or stealing model weights	Loss of IP; ability to run offline attacks on the stolen model

https://owasp.org/www-project-top-10-for-large-language-model-applications/

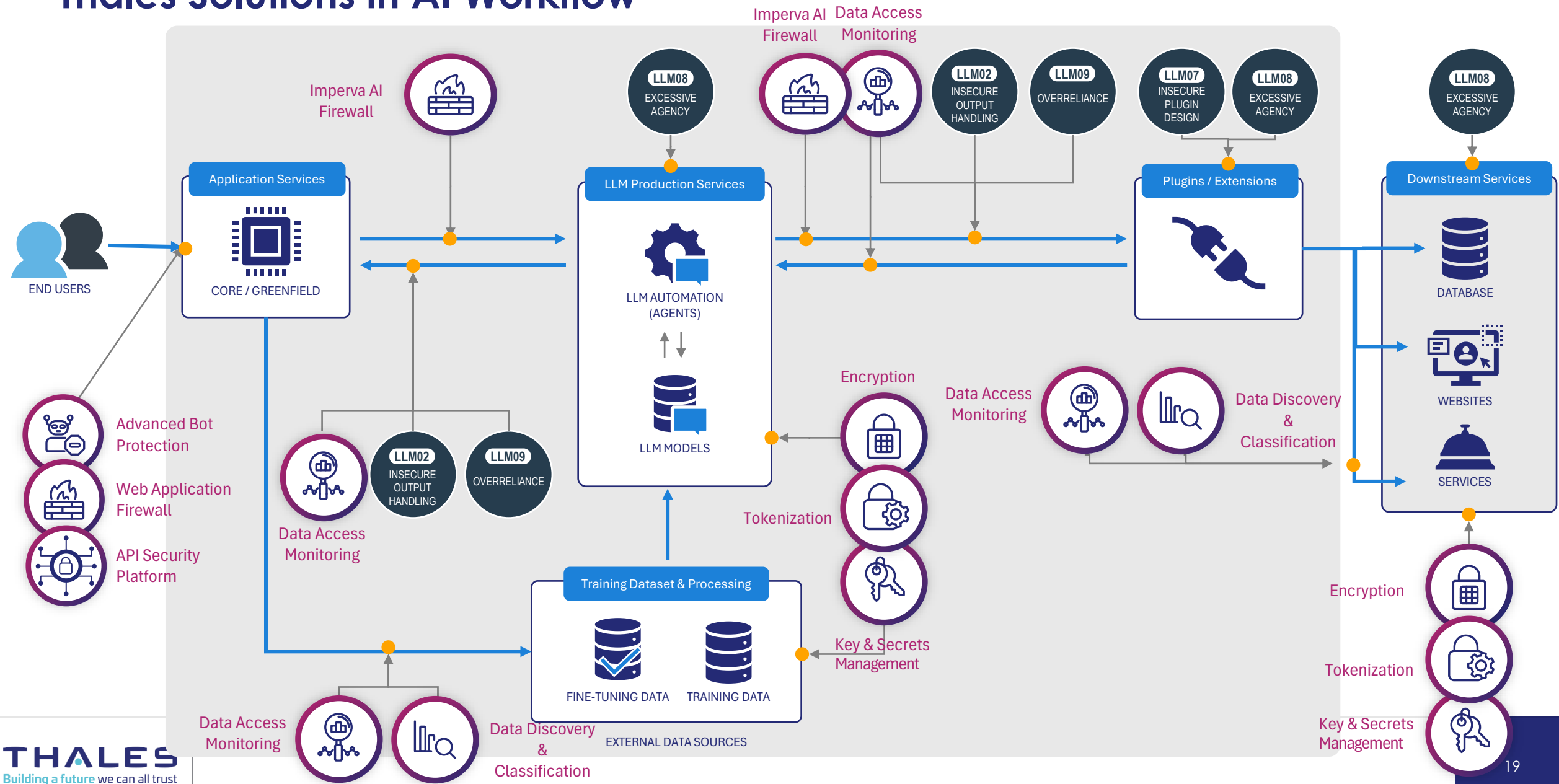
Revisiting a Common Enterprise GenAI or LLM Architectural Framework



OWASP GenAI Top 10 Visualized

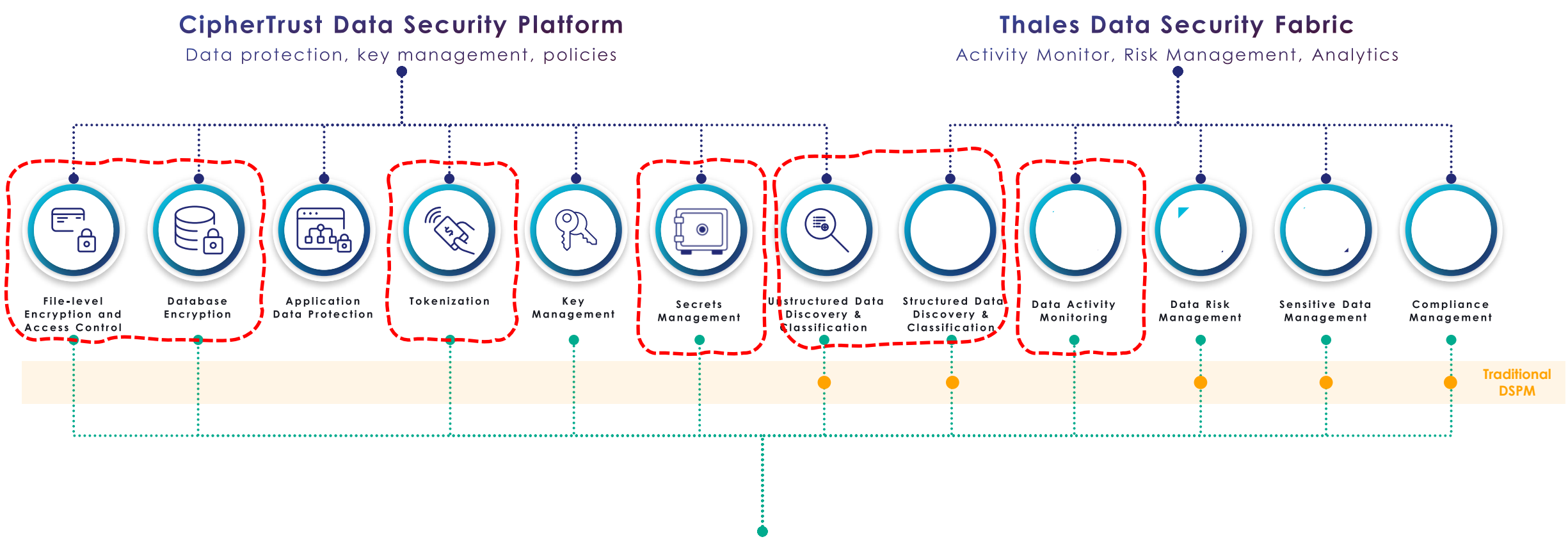


Thales Solutions in AI Workflow



Thales Data Security Platform

Jednolite bezpieczeństwo danych w środowiskach hybrydowych



Conclusion: Strengthen Your AI Security Posture with Enhanced Visibility, Control and Protection Against Threats to Data, Models and Applications

SECURE AI DATA

Protect Sensitive Data and Comply with Regulations in AI Initiatives

Data Discovery & Classification

Data Access Monitoring

Encryption

Tokenization

Key & Secrets Management

ENSURE INTEGRITY OF AI

Protect AI Applications and Models from Corruption, Misuse or Theft

API Security Platform

PROTECT AGAINST AI ATTACKS

Monitor Continuously for Risks and Threats from AI and to AI

Imperva AI Firewall

Web Application Firewall

Advanced Bot Protection

THALES
Building a future we can all trust

Thank you

...and AI ;-)

www.thalesgroup.com