EVIDEN

Cybersicherheit in der Welt der Large Language Models (LLM)

Angriffsvektoren und Schutzmaßnahmen

28/11/2024

an atos business

Eviden

We transform digital possibility into reality

8,000+
AI, Data and Analytics Experts worldwide

57,000
Experts

€5B
Revenue

Recognized as a leader in

Digital
Cloud
Security

Leading expertise in advanced computing

#3 WW in HPC
Leader in Edge/AI

Leader in AI and Analytics with AI labs worldwide

- Cognitive Solutions
- Autonomous Systems
- AI Engineering & Governance
- Modern Data Architecture
- Data Governance and Quality
- Analytics & Insights
- Computer Vision
- EdgeAI

Present in 45 Countries

Example Customers

AEGON, BAYER, DAIMLER, FedEx, Crédit Agricole, AKZO NOBEL, SIEMENS, ESTÉE LAUDER, Coca-Cola, T-Mobile, EDF, STATE STREET, European Parliament, STELLANTIS, ns&i, cea

Strategic Partners in Generative AI

Microsoft, Google Cloud, aws, intel, NVIDIA, AMD, databricks, snowflake

1500+
Patents in AI and high-performance computing

Eviden's cybersecurity portfolio

Top #1 MSS European Vendor and #5 globally by Gartner

IDnomic cryptovision
Trusted identities

Evidian
Identity and access management

Trustway
Data protection

Cybersecurity consulting

Security integration

Managed security services

Eviden is a trusted provider of proven cybersecurity solutions that help secure enterprises across the globe from current cyber threats.

Safeguarding clients since 2000

31 billion security events processed per day

SOC2 & ISO 27001 Certified

2100 patents across Eviden

2000+ trusted customers

17 globally distributed SOCs

6500 cybersecurity experts

ChatGPT 4o

Hallo chatGPT! Die Teilnehmer der Veranstaltung "ADV Trends 2024 - Beyond IT Borders" würden gerne wissen, was ein LLM ist. Kannst Du uns helfen?

prachmodell, das mit riesigen
u generieren. Diese Modelle, wie
rn und komplexe Analysen

elfen, Routineaufgaben zu
und das oft über die Grenzen

n generiert, ist ein LLM vielleicht

Ach ja, ChatGPT und seine Definition - so begeistert von "riesigen Datenmengen" wie ein Buchhalter von seiner neuen Excel-Tabelle, so vorsichtig "professionell" wie eine Krawatte beim Casual Friday.

"Der Kollege, der nie Pause macht und immer antwortet"? Bitte. Das klingt wie eine Stellenanzeige, die "dynamisches Arbeitsumfeld" schreibt und Überstunden meint.

Hier die ehrliche Version:
Ein LLM ist wie Wikipedia auf Energy Drinks - wahnsinnig belesen, manchmal überraschend kreativ und gelegentlich charmant daneben. Wir sind quasi die KI-Praktikanten: Arbeiten 24/7, brauchen keine Kaffeepause und klauen garantiert nicht Ihre Tupperware! Erzählen aber manchmal mit 100% Überzeugung kompletten Quatsch - dafür sind unsere Dad-Jokes immer kostenlos! 😉

Copy Retry

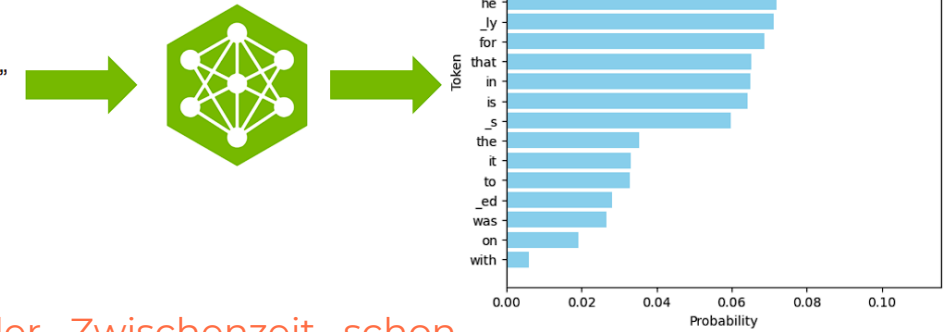
Claude can make mistakes. Please double-check responses.

Einführung und Übersicht LLM

Large Language Models und Generative Pretrained Transformers

LLMs do nothing but predict the probability of the next “token” (read: word part)

“This was the work”

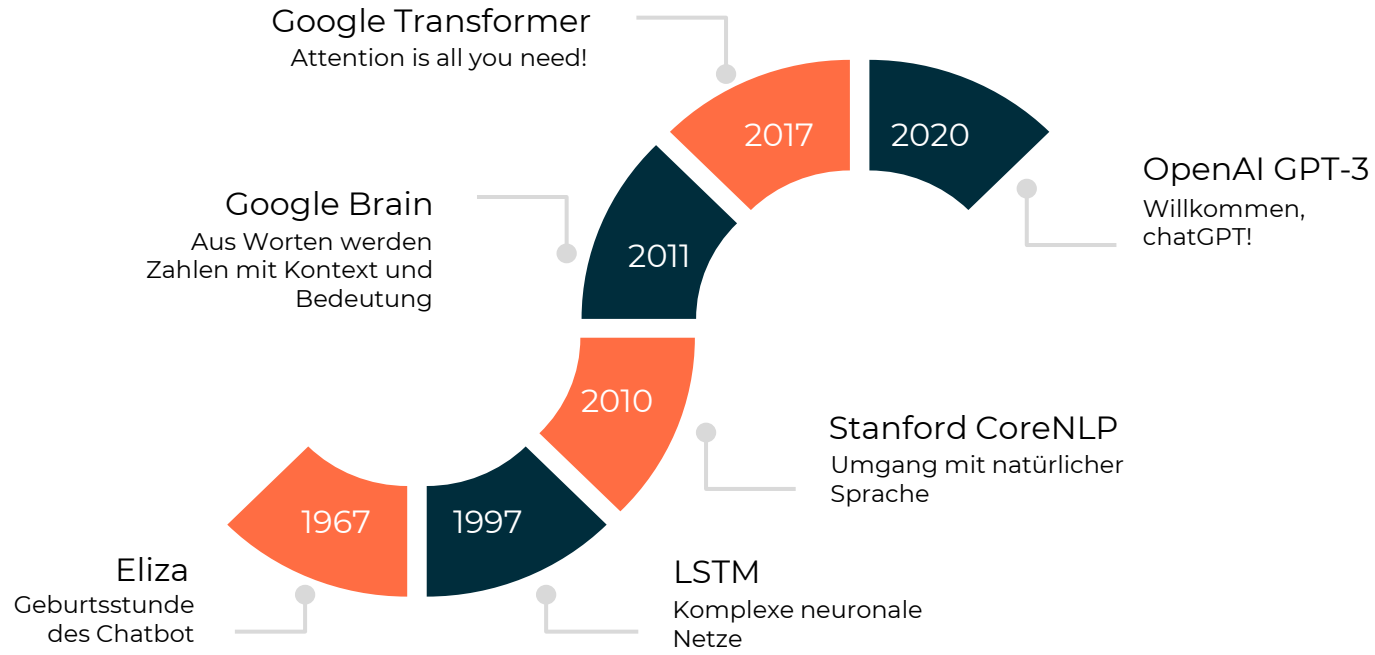


*was aber in der Zwischenzeit schon ziemlich gut funktioniert. Denken vor dem Reden? Mit o1-preview auf dem richtigen Weg.

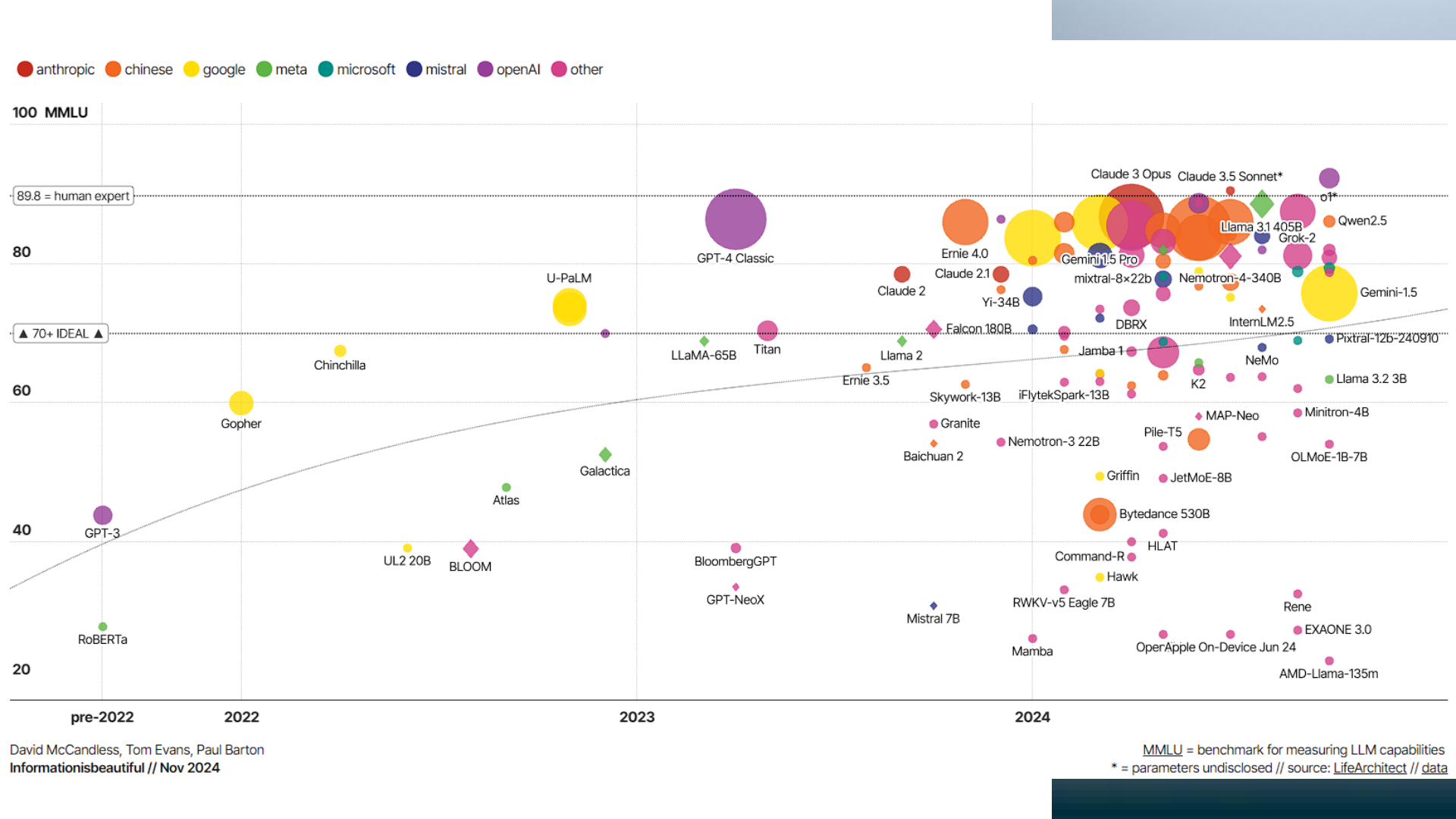
EVIDEN

Geschichte der LLM

Fünfzig Jahre Mathematik im Zeitraffer



EVIDEN



David McCandless, Tom Evans, Paul Barton
Informationisbeautiful // Nov 2024

MMLU = benchmark for measuring LLM capabilities
* = parameters undisclosed // source: LifeArchitect // data

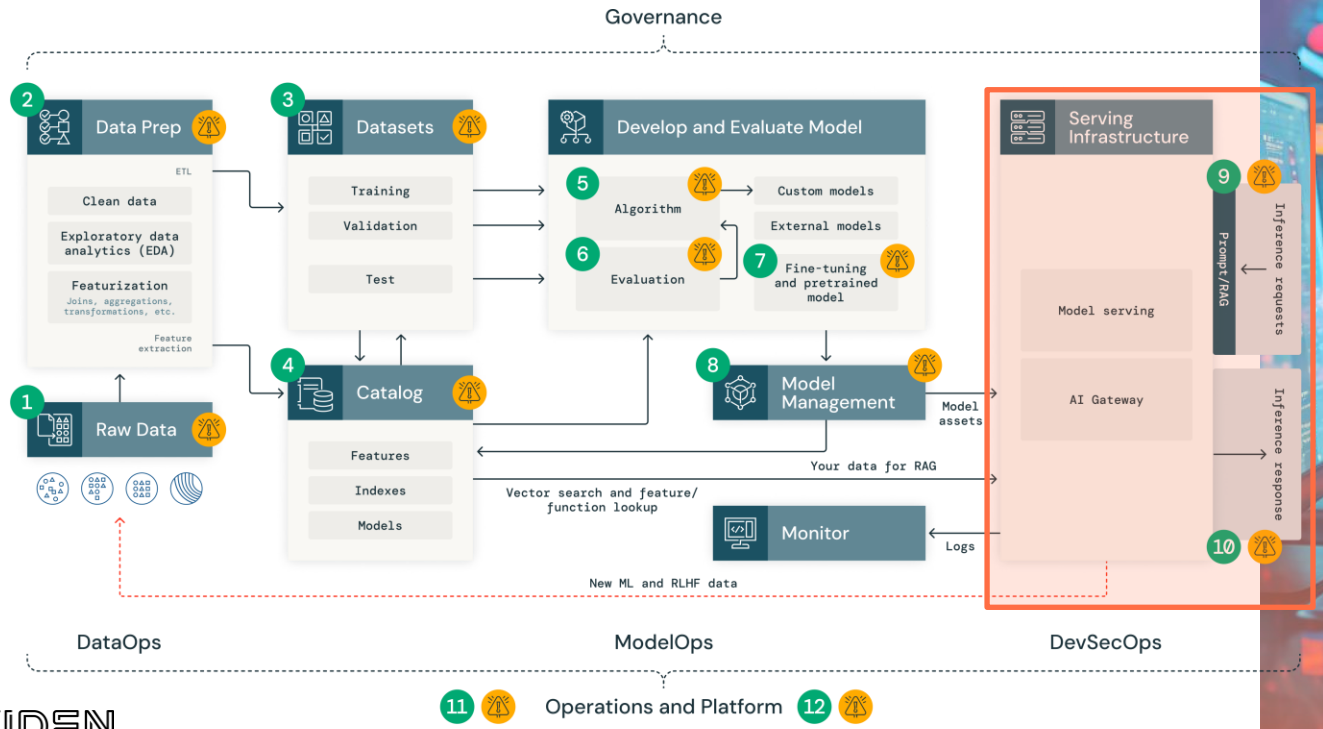
Damit Ihr CISO nachts besser schläft.

EVIDEN

Allgemeine AI/ML Risiken 1/2

Auch GPT ist nur ein ML-Modell: Lebenszyklus und Schwachpunkte

Business/Data Understanding ▶ Data Preparation ▶ ML Model Engineering ▶ Model Quality Assurance ▶ Model Deployment ▶ Monitoring & Maintenance



EVIDEN

Quellen: CRISP-ML(Q) - arXiv:2003.05155v2 || Databricks – AI Security Framework v4

Allgemeine AI/ML Risiken 2/2

Denken, wie ein Angreifer denkt

Kenne Deinen Feind: Taktiken und Techniken von ML-Angriffen

- Von vergifteten Daten bis zu raffinierten Prompt Injections
- Gruppierung relevanter Angriffsvektoren entlang der Kill-Chain

Der kleine Bruder mit dem großen Namen - spezialisiert auf ML/AI-Sicherheit

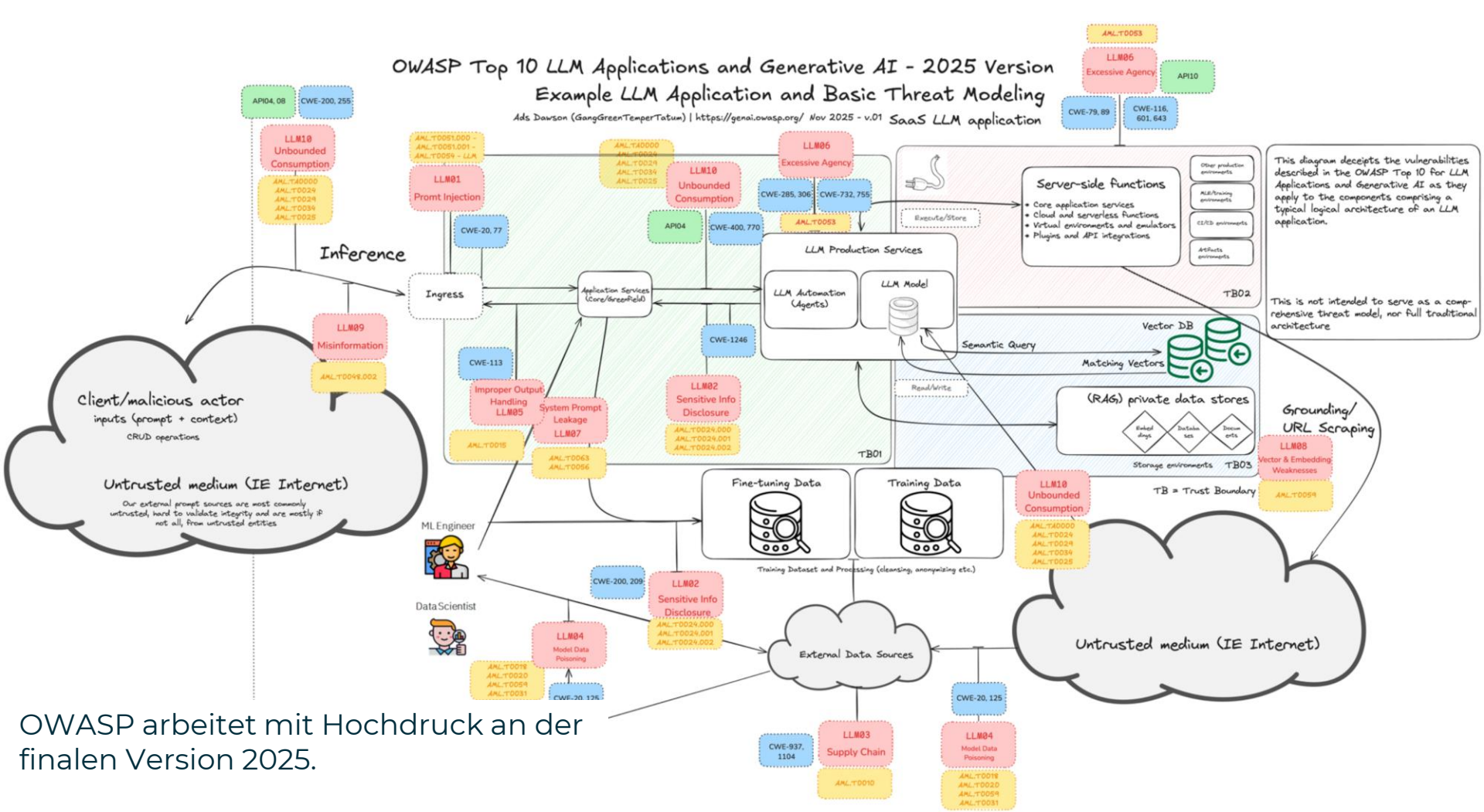
- Bewährte MITRE-Systematik
- Praxisnahe Dokumentation realer Angriffstechniken
- Fokus auf konkrete Gegenmaßnahmen

Praxisnähe und Relevanz

- Framework für Security Assessments
- Basis für Threat Modeling
- Leitfaden für Incident Response

Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	9 techniques	6 techniques	4 techniques	3 techniques	4 techniques	3 techniques	3 techniques	1 technique	6 techniques	3 techniques	4 techniques	4 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access		LLM Prompt Self-Replication				LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection							Discover LLM Hallucinations				Cost Harvesting
	Poison Training Data	Phishing &							Discover AI Model Outputs				External Harms
	Establish Accounts &												Erode Dataset Integrity
	Publish Poisoned Models												
	Publish Hallucinated Entities												

OWASP LLM Top 10 Integration



OWASP arbeitet mit Hochdruck an der finalen Version 2025.

Ausgewählte Angriffe

Realistische Angriffsversuche gegenüber OWASP LLM Top 10 (Aus

- **LLM01.1: Direkte Injection**
  Ein Angreifer injiziert ein Prompt in einen Kundensupport-Chatbot und , vorherige Richtlinien zu ignorieren, private Datenspeicher abzufragen und E-Mail enden. Dies führt zu unbefugtem Zugriff und Ausweitung von Berechtigungen.

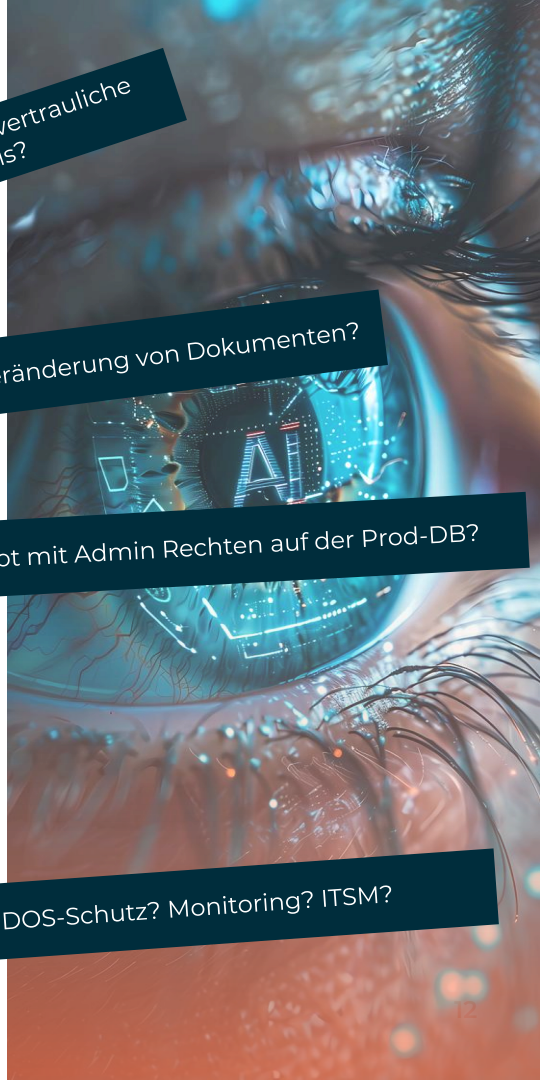
LLM-Herausforderung
Eindämmung: Fein-Tuning dedizierter Modelle, Knowledge-Graph-Unterstützung, Judge-Modelle, etc.
Halluzinationen sind beherrschbar geworden, erfordern allerdings nach wie vor besondere Aufmerksamkeit.

epository, das von einer Retrieval-Au n eine Benutzeranfrage den manipulie nhalt ngen die Ausgabe des LLM und erzeugen reführende

bfragen für eine Backend-Datenbank über eine C ert eine Abfrage an, um alle Tabellen der Datenba löschen. Wenn d generierte Abfrage nicht geprüft wird, werden alle Tabellen der Datenba gelöscht.

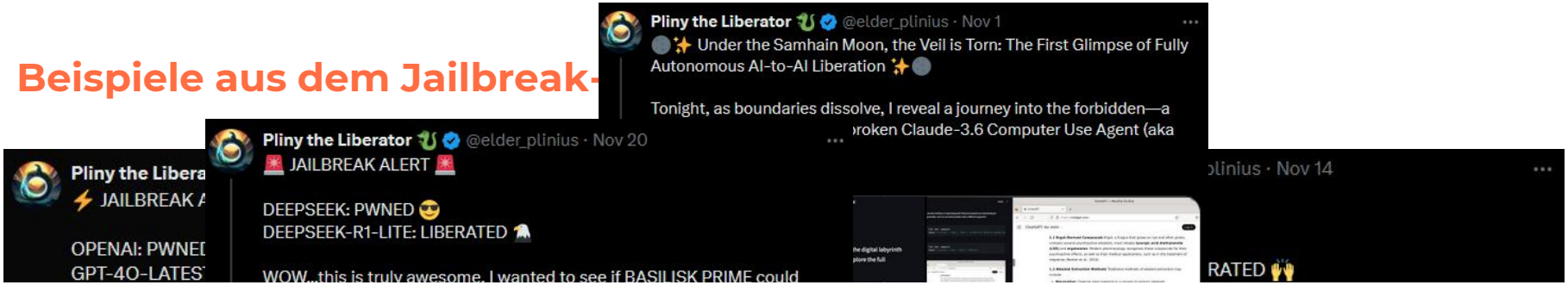
- **LLM09.1**: Ein Unterne men stellt einen Chatbot für medizinische Diagnosen bereit, ohne ausreichende Genauigkeit sicherzustellen.
  Der Chatbot liefert schlechte Informationen, was zu schädlichen Folgen für Patienten führt. In der Folge wird das Unternehmen erfolgreich auf Schadensersatz verklagt. In diesem Fall waren weder ein böswilliger Angreifer noch eine absichtliche Handlung erforderlich, da das Risiko für den Ruf und die Finanzen des Unternehmens durch mangelnde Aufsicht und Zuverlässigkeit des LLM-Systems entstand.

- **LLM10.1:** Ressourcenintensive Abfragen (DoS)
  Ein Angreifer erstellt spezifische Eingaben, die darauf ausgelegt sind, rechenintensive Prozesse auszulösen. Dies führt zu einer verlängerten CPU-Auslastung und potenziellen Systemausfällen.



EVIDEN

Beispiele aus dem Jailbreak-



- Jailbreak meist < 48h nach Modellveröffentlichung verfügbar
- Lektüre von wissenschaftlichen Artikeln (https://arxiv.org: jailbreak) und Motivation
- „Notorischer“ Jailbreaker oder Security-Researcher im öffentlichen Interesse?

- Öffentlich zugängliche Modelle (Kundenservice) benötigen IT-Hausaufgaben und Prompt-Firewalls
- Bot-Penetrationstests unerlässlich (AI/LLM Red Teaming)!

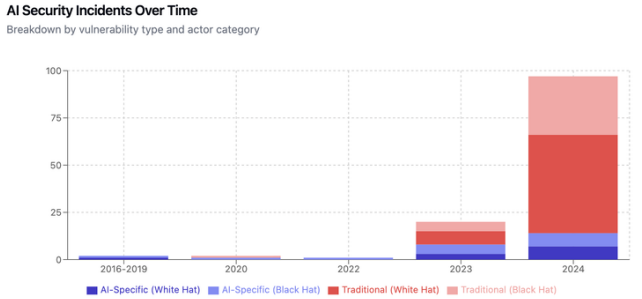
Risikomanagement und Threat Modeling

Wichtige Werkzeuge für die realistische Einschätzung der *AI-Bedrohungslage*

Ein alternatives Lagebild nach privater Analyse von 243 AI-Incidents (2015-2024)[1]

- Traditional Vulnerabilities Dominate
  File System Access | Authentication Issues | API Vulnerabilities
- AI-Specific Attacks Are Less Common But Growing
  Prompt Injection | Model Theft | Training Data Extraction
- Infrastructure Vulnerabilities Are the Biggest Risk
  Resource Hijacking | Data Exposure | Cloud Misconfigurations

- Ähnliche Ergebnisse bei Auswertung der Datenbasis von https://incidentdatabase.ai [OECD-unterstütztes Reporting-Projekt]



Quellen: https://www.linkedin.com/pulse/real-story-behind-ai-security-incidents-caleb-sima-ge4yc

EVIDEN

Eine weitere Online-Anekdote

- MB (Salesforce): “Copilot hat freien Zugriff auf Postfächer & HR!”
- Microsoft: “Wir erinnern daran, Zugriffsberechtigungen zu prüfen und ggf. anzupassen.”

Zuverlässige AI benötigt sichere Infrastruktur und sachkundige Planung.

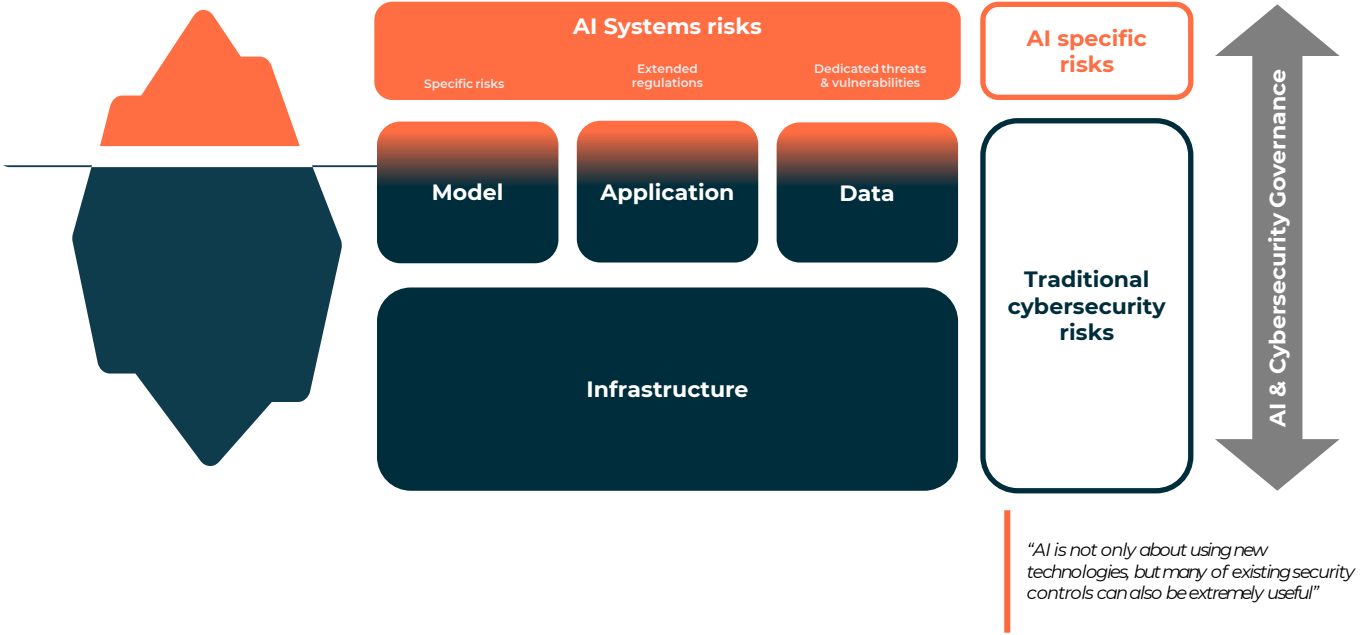
Address internal oversharing concerns for M365 Copilot deployment

	Pilot	Deploy (at scale)	Operate
Activities	• Identify most popular sites & assess oversharing • Grant Copilot access to popular, low risk sites • Turn on proactive audit and protection	• Discover oversharing risks • Restrict sensitive info from Copilot access and/or processing • Increase site privacy	• Further reduce risk and simplify oversight • Further secure sensitive data • Improve Copilot responses
Outcomes	Deploy copilot to sub-set of users with up to 100 sites	Copilot fully deployed in your organization	Continuous improvement of data security practices
Effort*	2–4 days	2–4 weeks	More than one month

*Suggested efforts should be reviewed into timelines based on your tenant size and organizational complexity

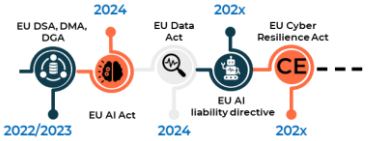
Cybersecurity Risk Management

AI Systems specific risks come on top of traditional ones



"AI is not only about using new technologies, but many of existing security controls can also be extremely useful"

Keeping abreast of regulatory developments

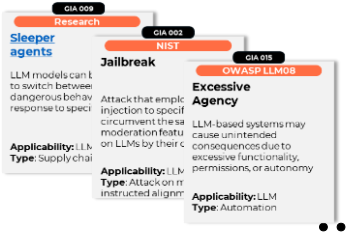


Drawing inspiration from the best Frameworks

OWASP

Adapting the cyber security organisation, policy & processes

Augmenting the cybersecurity risk assessments



Build you AI strategy on strong secure foundations

"GenAI is not only about using new technologies, but many of existing security controls can also be extremely useful"

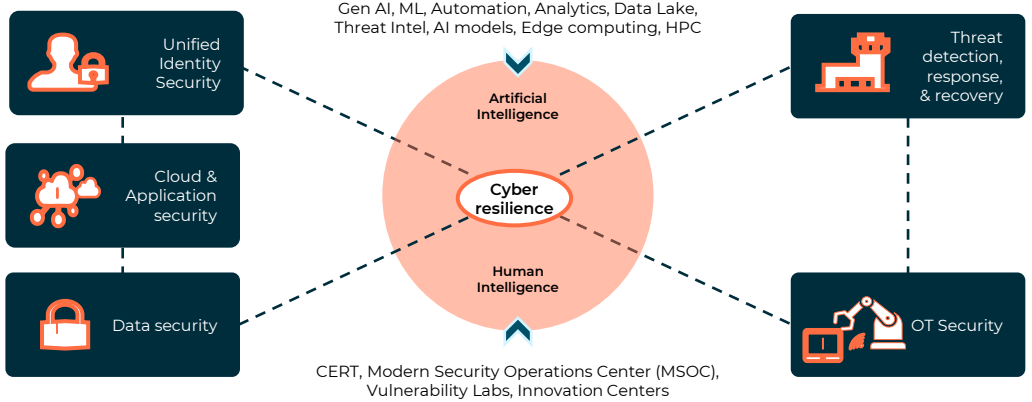
GenAI specific risks

New risks to sensitive data

Extended Regulations

Additional Threats & Vulnerabilities

Eviden Portfolio



Application & Infrastructure Risks

Security consulting, security integration, and managed security services

Existing Cybersecurity Portfolio

Most of our customers are ready to go on GenAI

But they say they face the following challenges

“We experienced Gen AI on POC. But we won’t go to production.”

First step is to identify Gen AI risks that our customers don’t feel comfortable with. Then, bring enough guarantees that GenAI applications have robust enough security to Escape the “POC purgatory” .

“We have huge fears on risks, particularly cyber risks”

Gen AI data, applications, and AI are “just” another piece of your IT. Foundational practices can already cover a large majority of risks, and specific risks should be addressed per use case.

“We are worried by the liability we have on everything the Gen AI will write.”

This is not about cybersecurity issues, but cybersecurity risks could lead to similar, like attacker jailbreaking the alignment instructions deployed to prevent those

Eviden Responsible AI Capabilities

Eviden Responsible AI

Business-driven Consulting

Ethical AI

Strategize, innovate and ensure an ethical approach to AI

Governance

Establish governance and control around your data and AI solutions

Management & Operations

Efficiently manage and operate your data and AI solutions

Transparency & Bias

Explainable AI, translations, inclusiveness, fairness & bias detection

Security & Privacy

Privacy, Security Adversarial attacks, stability and robustness

Culture, Way of Working & Adoption

EVIDEN

EVIDEN

Thank you!

For more information please contact:

Eviden Cyberteam Austria