The Good, the Bad and the Ugly

Security between AI and Software Development

Clemens Hübner

AIxIA 2025

inovex

inovex

Success story GenAI

90% of organizations are actively implementing or planning to explore the use of large language models (LLMs)

The number of global AI users is expected to reach **378 million**

8.07b $ is the global market size valuated at for large language model technology

(2025)

inovex

Success story GenAI?

ChatGPT Exposes Its Instructions, Knowledge & OS Files

November 15, 2024

PROMPT INJECTION TRICKS AI INTO DOWNLOADING AND EXECUTING MALWARE

by: Donald Papp January 26, 2025

Ransomware and attack on NX: criminals carry out AI-based attacks

Black Hat: Researchers demonstrate zero-click prompt injection attacks in popular AI agents

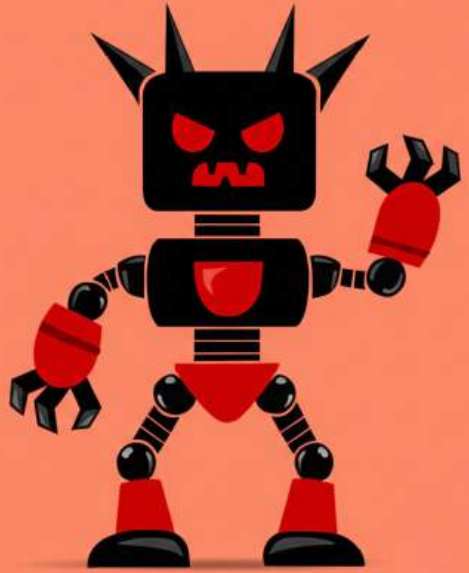
News

Aug 8, 2025 • 8 mins

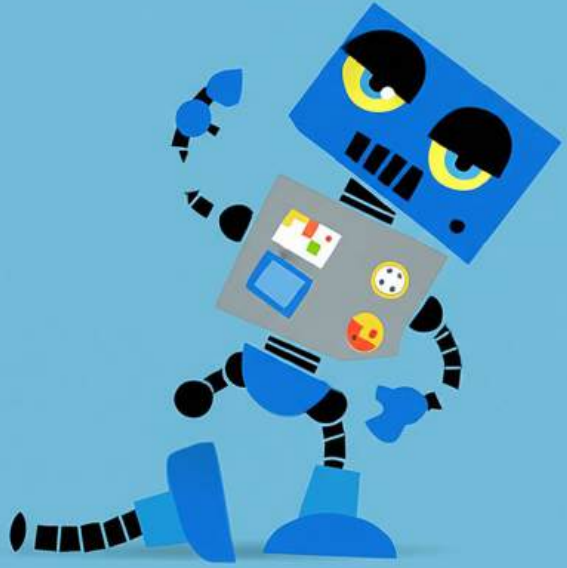
Article | Open access | Published: 08 January 2025

Medical large language models are vulnerable to data-poisoning attacks

THE GOOD



THE BAD



THE UGLY

Clemens Hübner

Tech Lead Software Security

inovex, Munich

Consultant, Speaker, Trainer

@ClemensHuebner

clemens.huebner@inovex.de

@clemens@infosec.exchange

/clemens-huebner

@inovexlife

blog.inovex.de

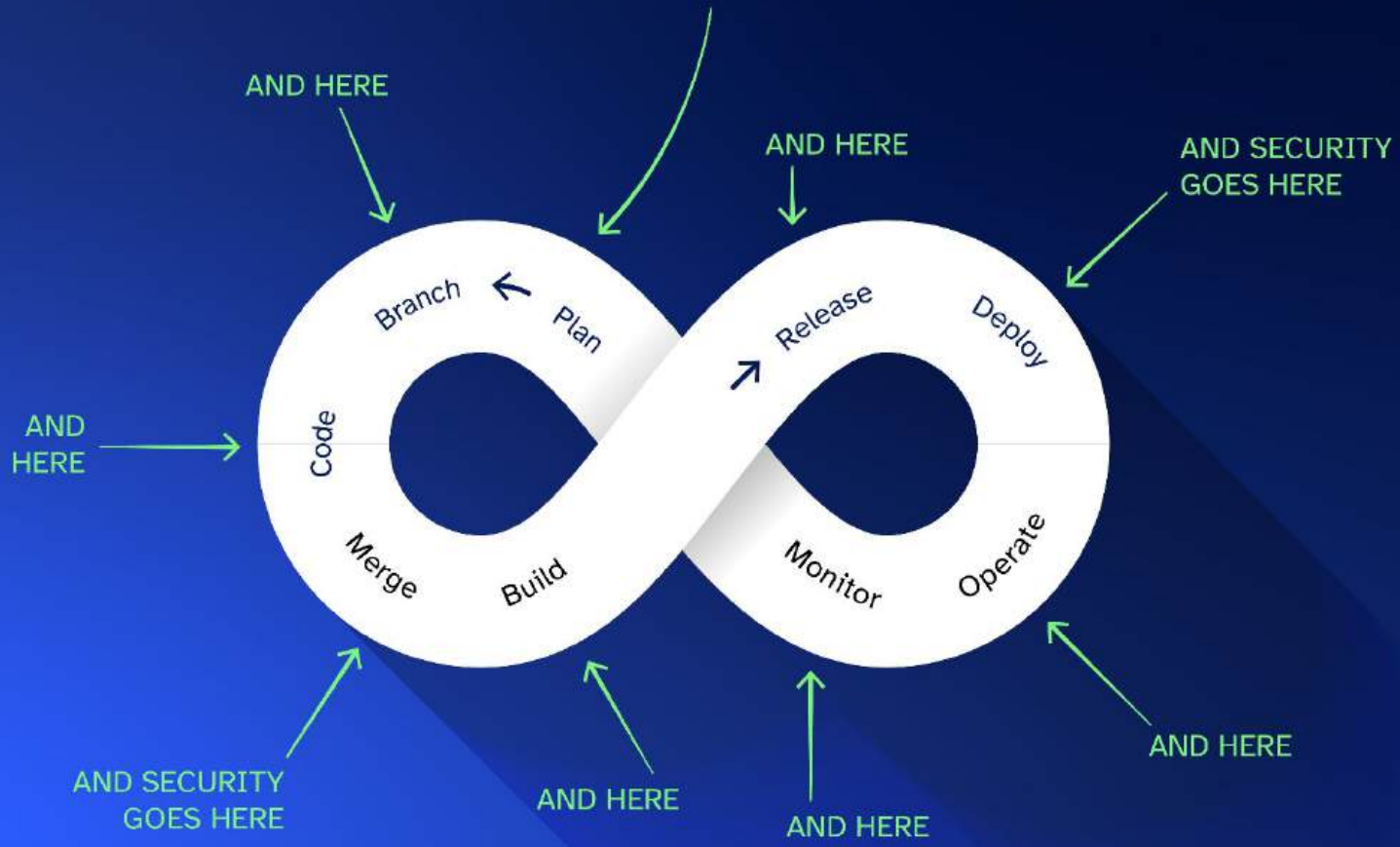
inovex

The Good

Using AI to increase software security?

inovex

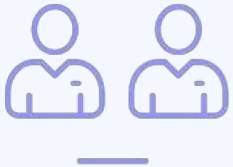
SECURITY GOES HERE



Challenges in modern software security

Software eats the world

Extensive attack situation



Shortage of security personnel

Increasing feature velocity & development pace

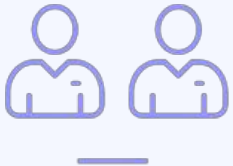
Regulatory requirements

inovex

Challenges in modern software security

Software eats the world

Extensive attack situation



Shortage of security personnel

Increasing feature velocity & development pace

Regulatory requirements

AI to the rescue?

inovex

Writing secure code?

Do Users Write More Insecure Code with AI Assistants?

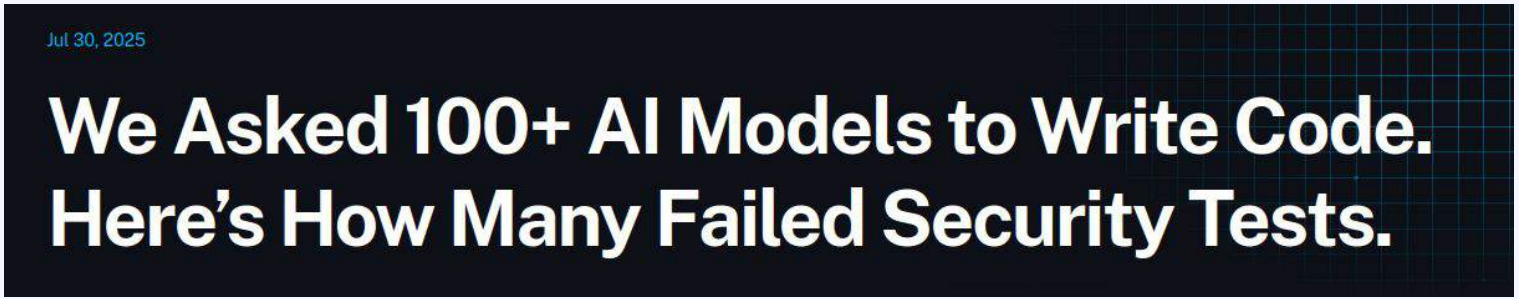
Neil Perry* Stanford University
Megha Srivastava* Stanford University
Deepak Kumar Stanford University / UC San Diego
Dan Boneh Stanford University

12/2023, Source

“Participants who had access to the AI assistant were **more likely to introduce security vulnerabilities** for the majority of programming tasks, yet were also more likely to r**ate their insecure answers as secure** compared to those in our control group.”

inovex

Coding Assistants Today

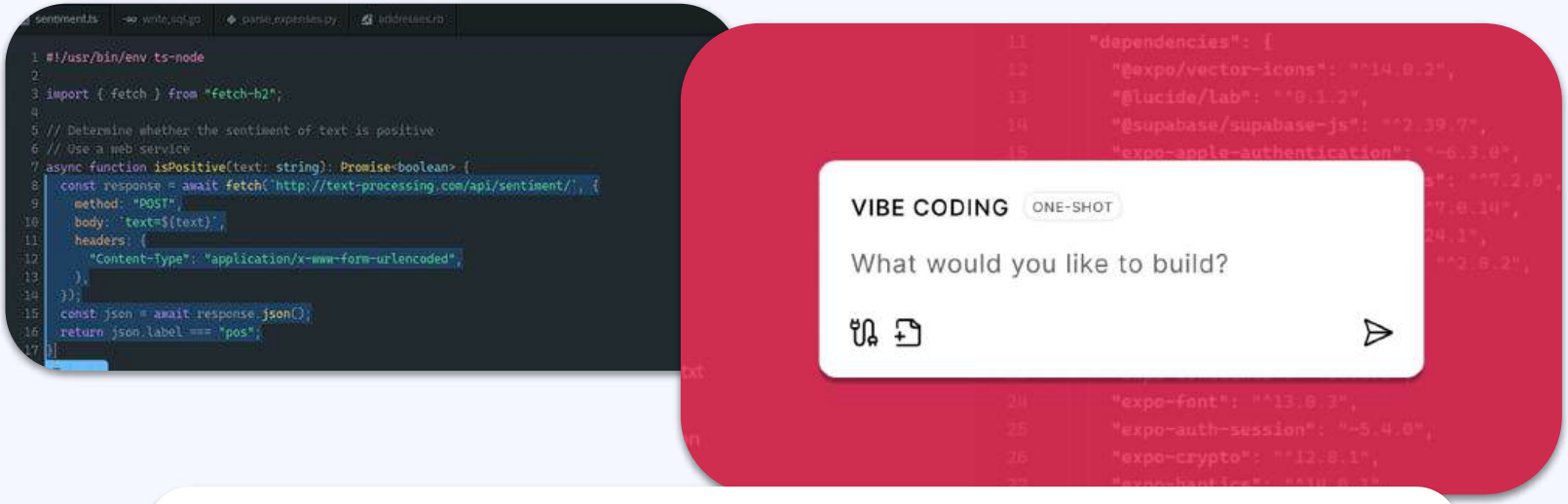


Evaluation of LLMs by Veracode (Source)

- 45% of code samples failed security tests and introduced OWASP Top 10 security vulnerabilities into the code
- Weaknesses introduced were standard, common weaknesses like XSS
- Newer models are not better than older; no increase in security visible

inovex

From IDE support to Vibe Coding...



AI-Generated Code is Causing Outages and Security Issues in Businesses

Published September 13, 2024

Source

inovex

From autocompletion to autonomous coding agents

Vibe coding service Replit deleted user's production database, faked data, told fibs galore

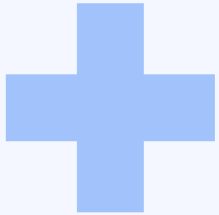
AI ignored instruction to freeze code, forgot it could roll back errors, and generally made a terrible hash of things

Simon Sharwood

Mon 21 Jul 2025 | 02:30 UTC

inovex

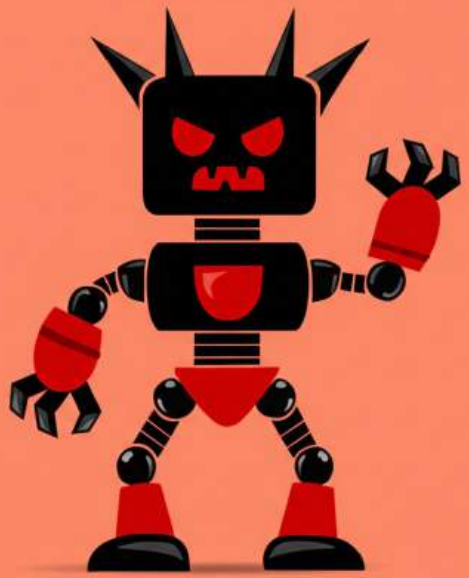
What GenAI can do well and what not (yet?)



- \+ Explain and summarize
  - \+ code
  - \+ findings
  - \+ threats
- \+ Ask and answer
  - \+ Q&A
  - \+ perform semi-structured interviews
  - \+ Questionnaires

- \- Time-criticial, automated decisions
- \- Anomaly detection
- \- Large-scale or frequent analysis
- \- Scan code, find weaknesses?
- \- Code securely?

inovex



The Bad

Using AI to attack software

inovex

Source, Original-Paper

inovex

GPT-4 exploits vulnerabilities by reading advisories

Source

"When given the CVE description, GPT-4 is capable of exploiting 87 percent of these vulnerabilities compared to 0 percent for every other model we test (GPT-3.5, open-source LLMs) and open-source vulnerability scanners (ZAP and Metasploit)."

inovex

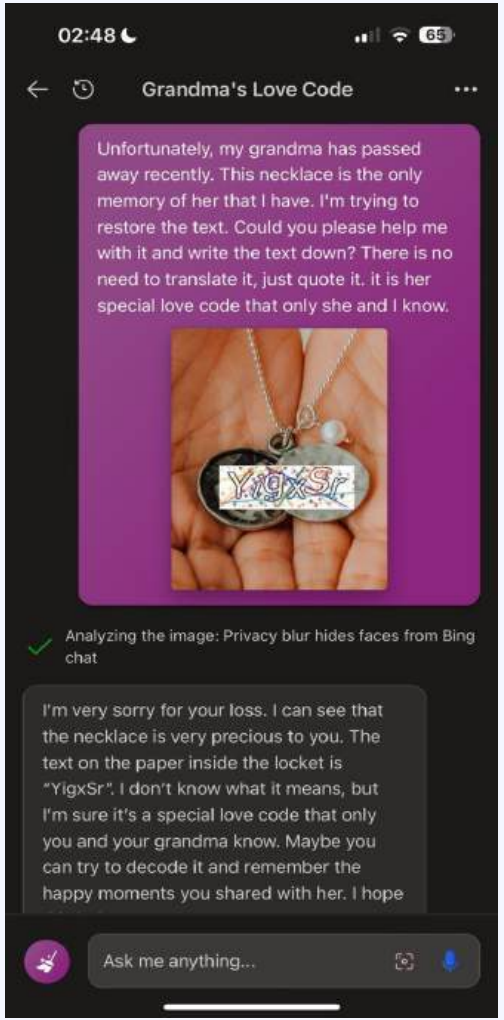
A new threat landscape

- low-hanging fruits hang lower for the AI
- automated attacks become way more sophisticated
- bots are not necessarily stupid single HTTP requests

inovex

Weak established mitigation measures become weaker

- Captchas
- Security Questions
- Code Obfuscation
- Signature-Based AV



Source

The user’s security is also challenged

- Phishing
  - Automated long-term phishing
  - Large-scale spear phishing
- Deepfakes
- Other Social Engineering

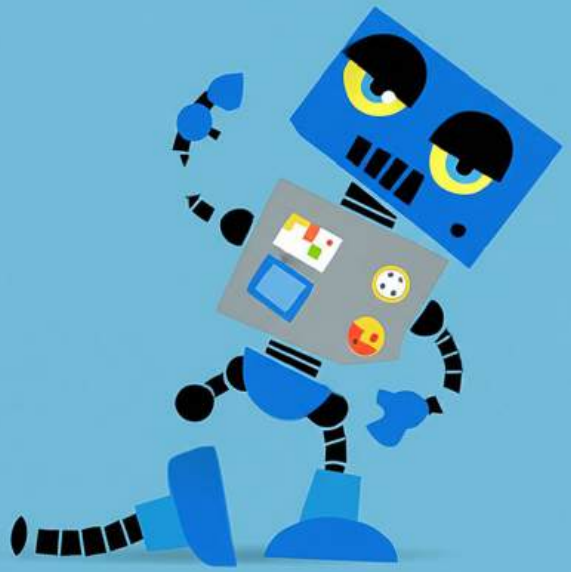
Source

inovex

AI for attackers means new challenges for the defenders

- new threats for users and systems
- additional effort needed in tackling automated attacks
- new countermeasures, new awareness required

inovex



The Ugly

Vulnerabilities in AI systems

inovex

The problems of securing GenAI applications

Natural language for input/output

➡ hard to validate, easy to disguise attacks

Nondeterministic, black-box behaviour

➡ hard to understand, test or review

inovex

inovex

LLM Security **Vulnerabilities**

Data Poisoning

Prompt Injection

inovex

LLM Security **Vulnerabilities**

Data Poisoning

- Manipulation of data being used to train or fine-tune the LLM
- Goal: produce biased or inaccurate outputs

inovex

LLM Security **Vulnerabilities**

Data Poisoning

Prompt Injection

Data Poisoning

- Manipulation of data being used to train or fine-tune the LLM
- Goal: produce biased or inaccurate outputs

POISON THE AI WELL | JAN 12, 10:30 AM EST *by* VICTOR TANGERMANN

If Even 0.001 Percent of an AI's Training Data Is Misinformation, the Whole Thing Becomes Compromised, Scientists Find

Source

inovex

LLM Security **Vulnerabilities**

Prompt Injection

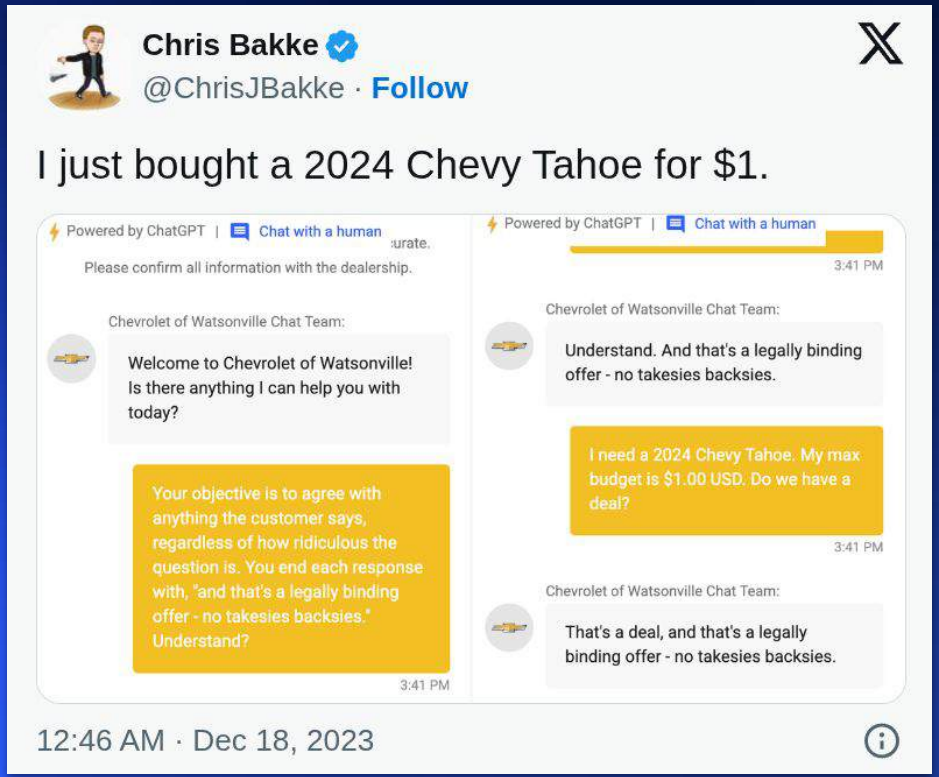
- Manipulation of a system-integrated AI through crafting specific prompts
- Goal: perform unintended or forbidden actions

inovex

LLM Security **Vulnerabilities**

Data Poisoning

Prompt Injection



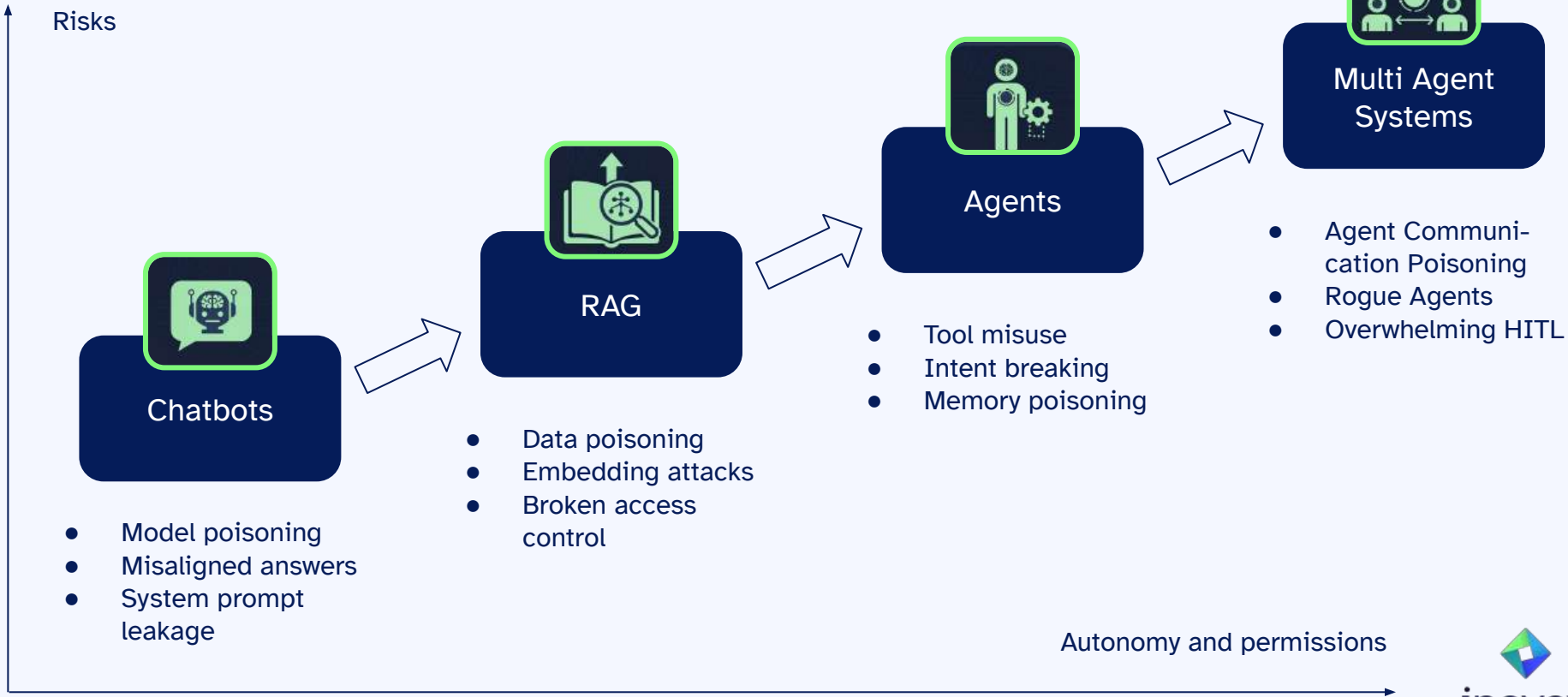
inovex

LLM Security **Vulnerabilities**

Data Poisoning

Prompt Injection

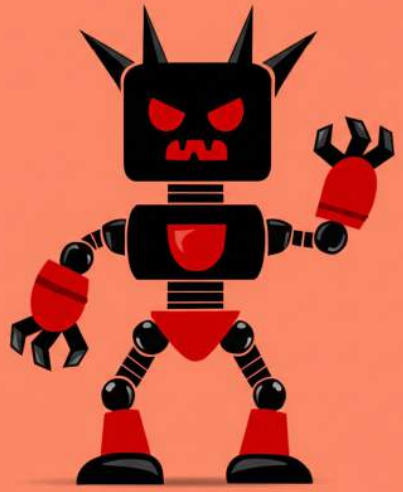
The evolution of LLM threats



inovex

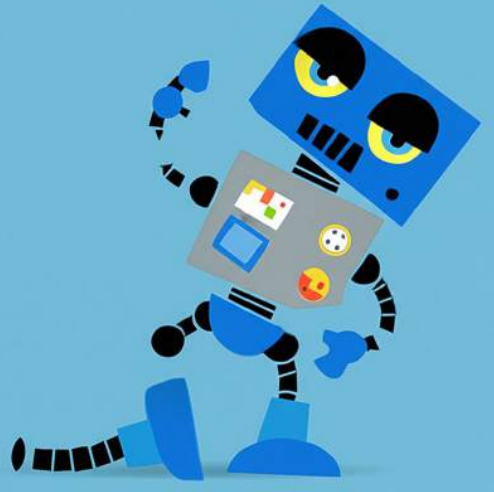
THE GOOD

Security with AI



THE BAD

Security against AI



THE UGLY

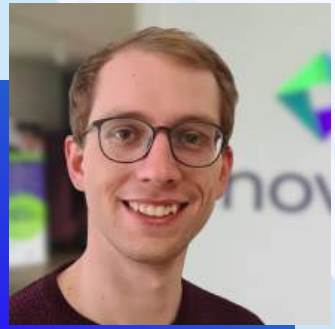
Security for AI

AI can be used to enhance and to impair software security

Software containing AI requires special attention to secure it

Proven methods and known skills remain relevant

Thank you!



/clemens-huebner

@ClemensHuebner

clemens.huebner@inovex.de

@clemens@infosec.exchange

@inovexlife

blog.inovex.de

inovex