GEDOPLAN

LLMs in der Praxis mit Jakarta EE

Anbindung mit LangChain4j

GEDOPLAN GmbH

Jan Pohlmeyer

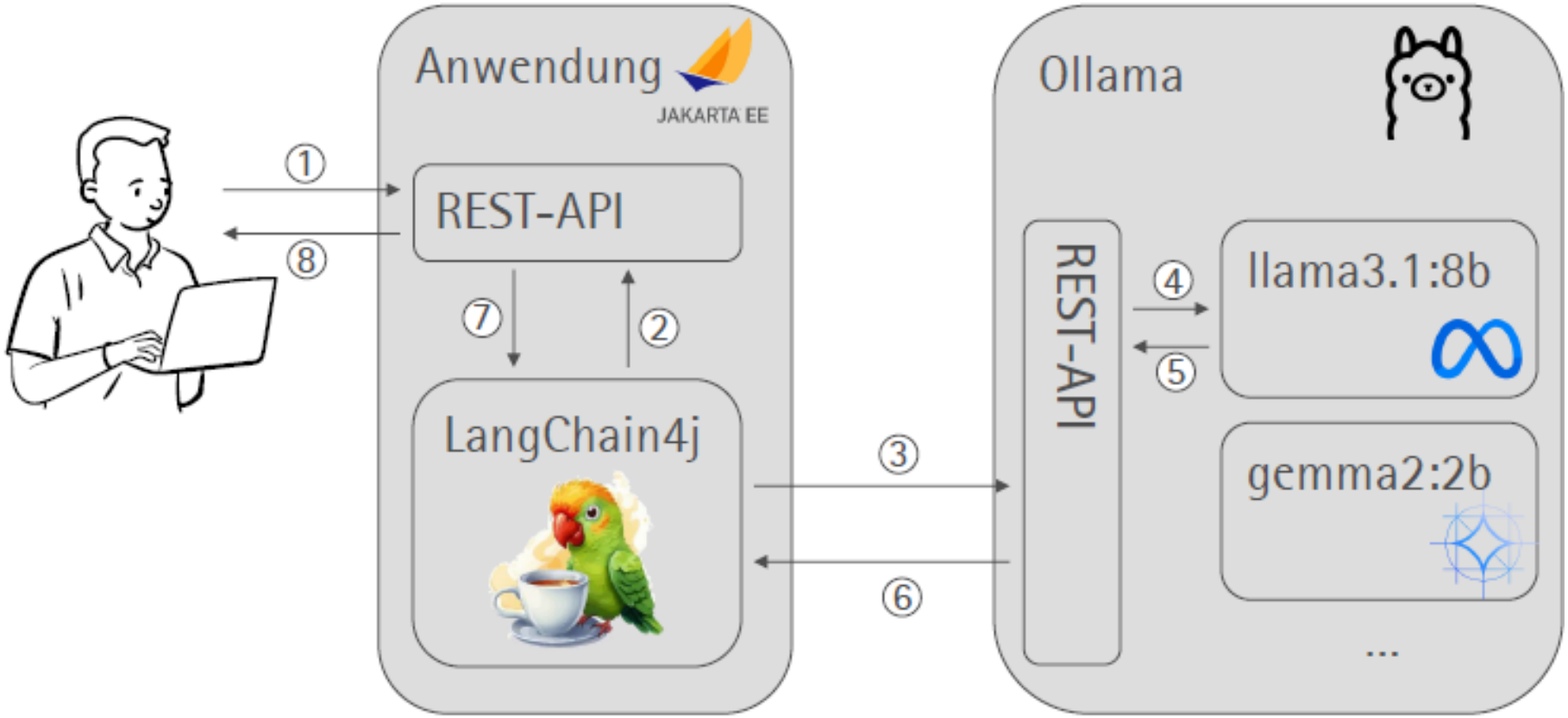
LangChain4j

- Bibliothek zur Abstraktion von KI-Inferenzschnittstellen
- Stammt nicht direkt von Python/JS Bibliothek LangChain ab
- Open-Source Entwicklung (Apache 2.0 Lizenz)
- Initial hauptsächlich für LLMs
  - mittlerweile auch weitere Schnittstellen z.B. für Bildgenerierung
- Minimum JDK 17
- Weiteres:
  - https://docs.langchain4j.dev
  - https://github.com/langchain4j/langchain4j

Ollama

- lokale "Ausführungsumgebung" für LLMs
- Open-Source Entwicklung (MIT Lizenz)
- Einfaches Setup für diverse Systeme und Architekturen
- REST-API zur Kommunikation mit einem ausgeführten LLM
- Weiteres:
  - https://ollama.com/
  - https://github.com/ollama/ollama

Beispiel Kommunikationsweg

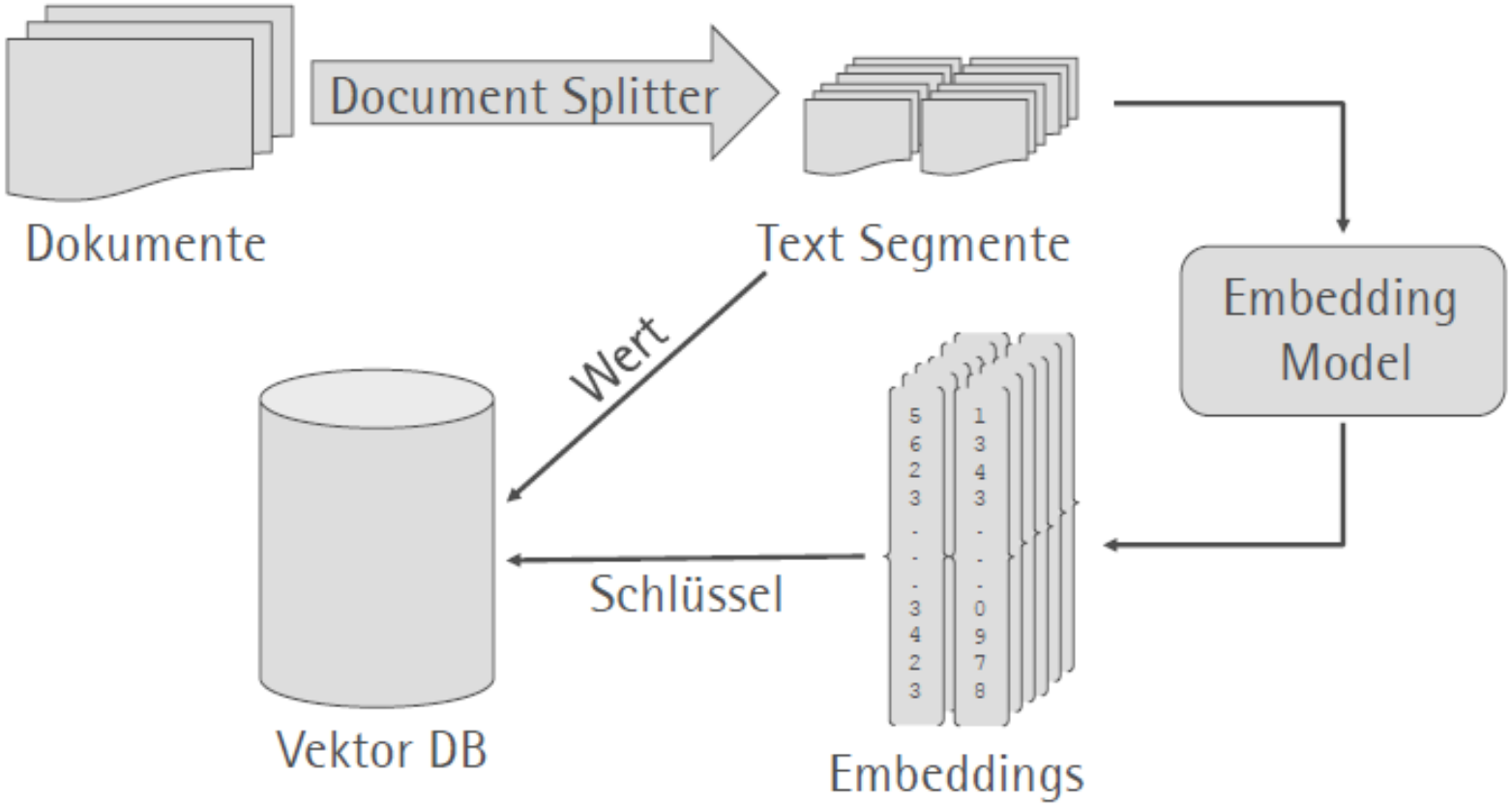


Demo

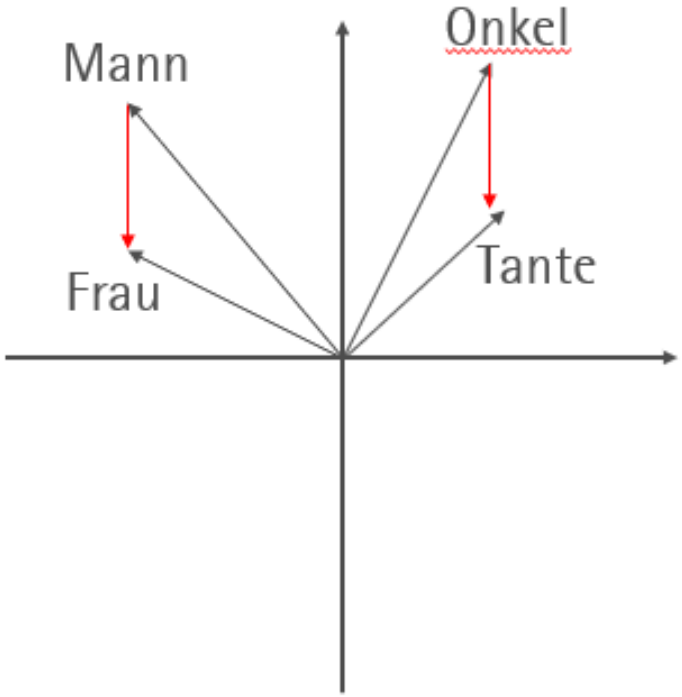
- ChatLanguageModel (+MPConfig)
- SimpleAiService
  - HTTP
  - Plain HTML/JS
  - Jakarta Faces
- MemoryAiService

- Github: https://github.com/GEDOPLAN/jee-langchain4j-ekj-demo

RAG mit Embeddings: Ingestion

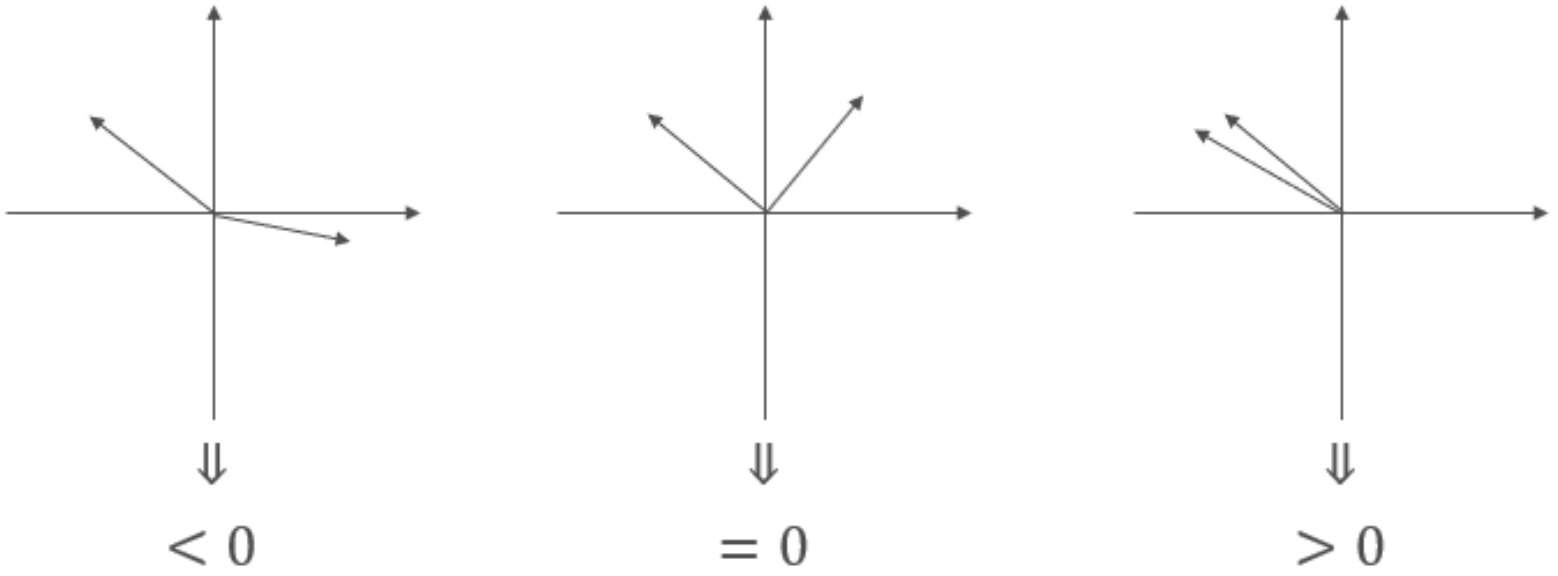


Embeddings

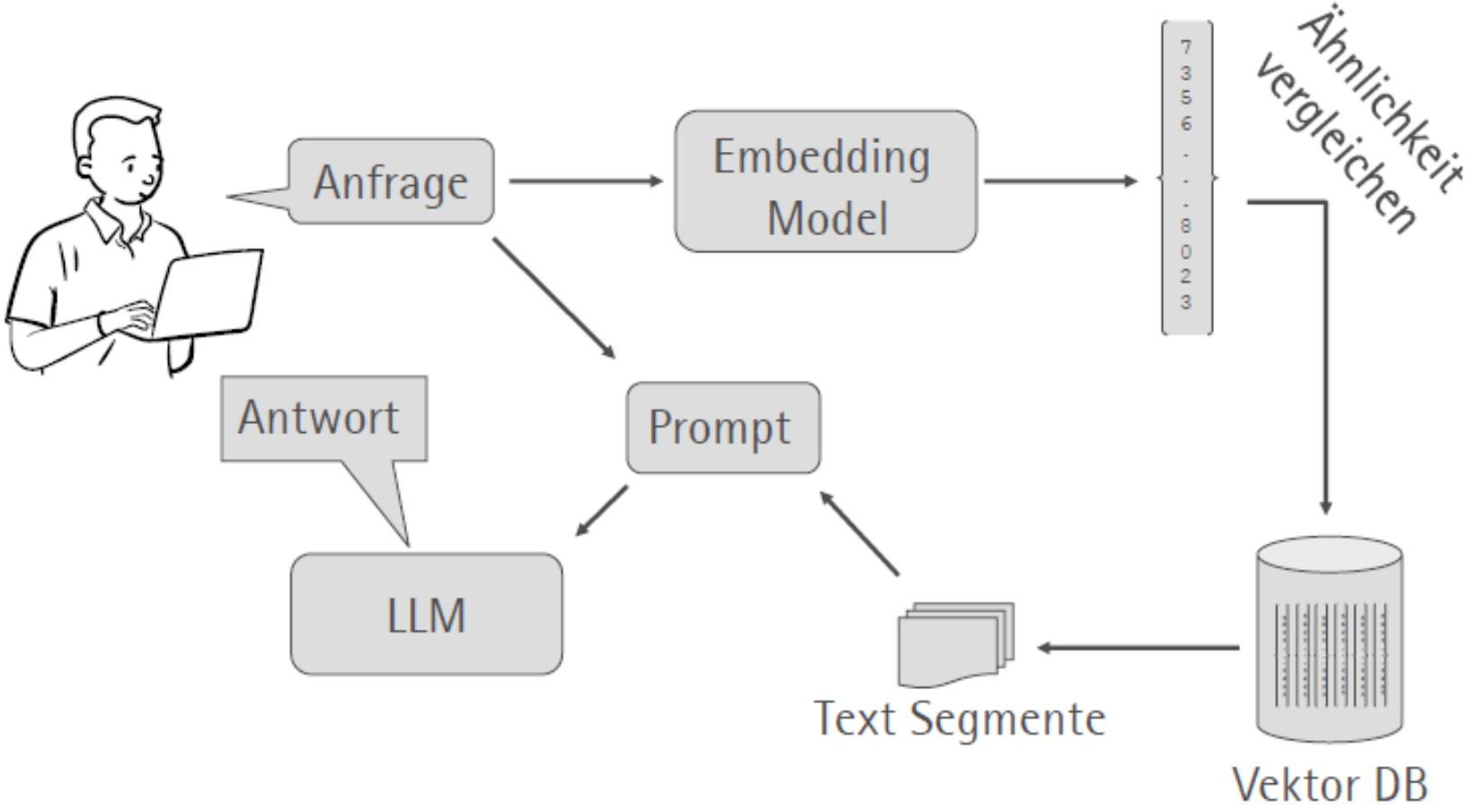


Ähnlichkeit von Embeddings (Bedeutung) messen

- Ähnlichkeit der Richtung ⇒ Kosinus-Ähnlichkeit



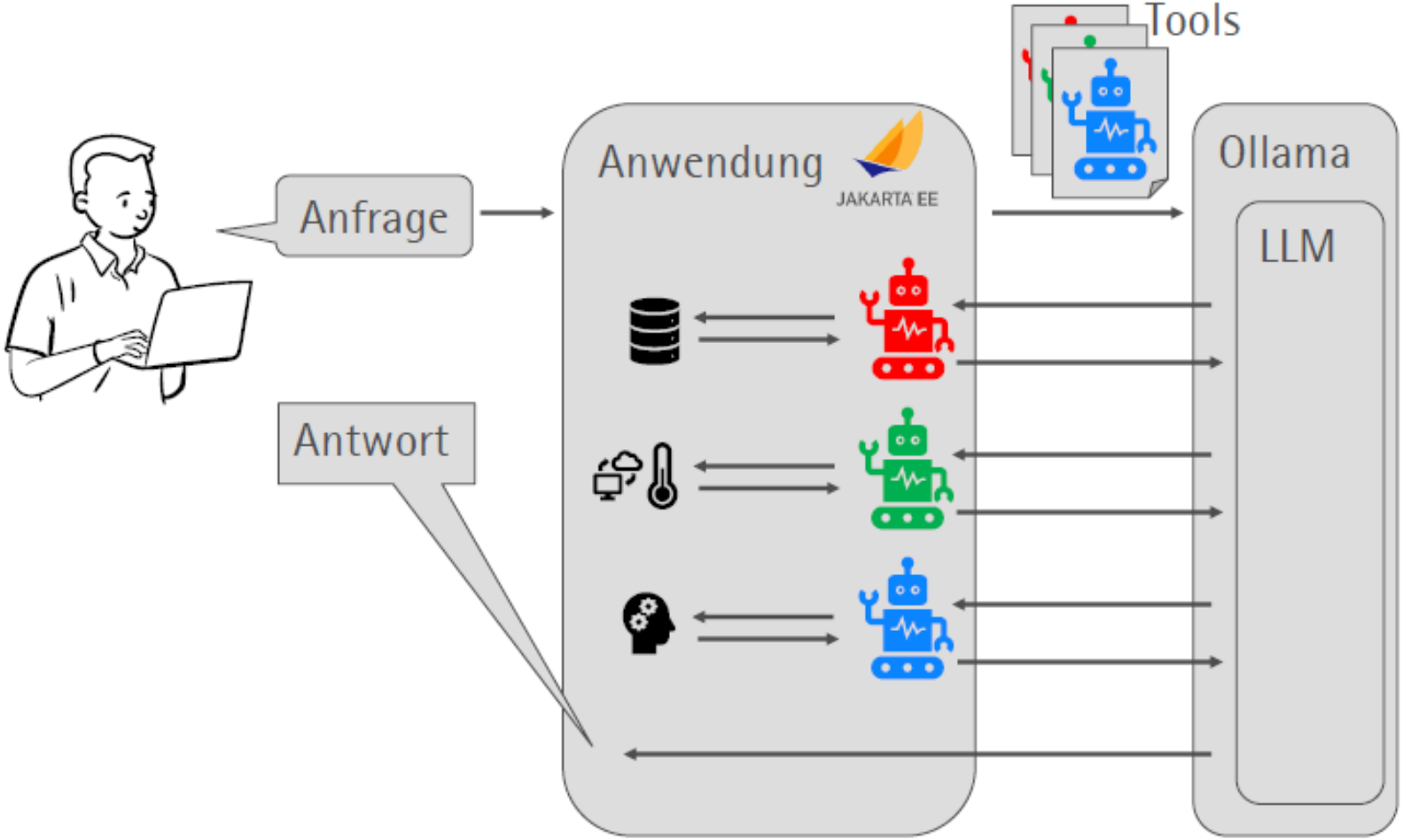
RAG mit Embeddings: Retrieval



Demo

- RagAiService

Mit Tools werden die LLMs zu Agenten



GEDOPLAN

Demo

- ToolsAiService

Chancen/Möglichkeiten

- Textverarbeitung jeglicher Art
  - Formulierungshilfen
  - Inhaltliche Zusammenfassungen
  - Brainstorming
  - Programmierhilfen
  - Erstellen von Dokumentation
  - Übersetzung
  - Natürlichsprachliche Interaktion mit technischen Systemen
  - ...

Probleme und Risiken

- Nichtdeterminismus "by design"
  - Nicht alles was heute funktioniert, klappt auch morgen
  - LLMs können (objektiv betrachtet) Fehler machen
    - Auf Ergebnis kann man sich nicht zu 100% verlassen
  - Stochastische Maschine ohne formale Logik (nur Approximation)
- Halluzinationen oder Bullshitting
- Prompt Injection
- Excessive Agency
- System Prompt Leakage
- ...
- OWASP LLM Top 10: https://genai.owasp.org/llm-top-10/

Noch Fragen?

- Nächster Expertenkreis am 08.05.2025