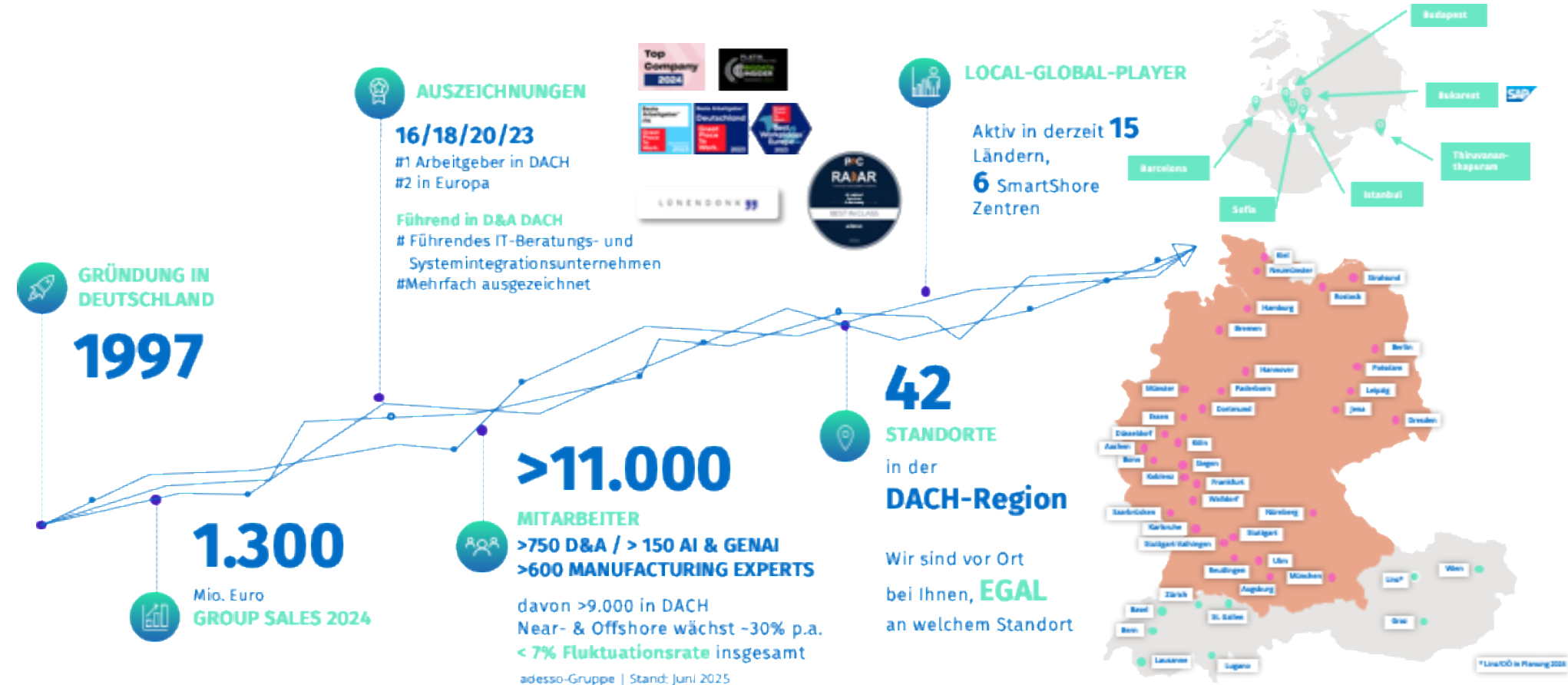






Jan Skrovanek

- Seit Oktober 2024 bei adesso im Team „GenAI Solutioning Unit“
- Vorher 2,5 Jahre bei Google
- Zuständig für die Geschäftsentwicklung fürs Thema Generative KI in der öffentlichen Verwaltung, dem Maschinenbau sowie der Automobilindustrie

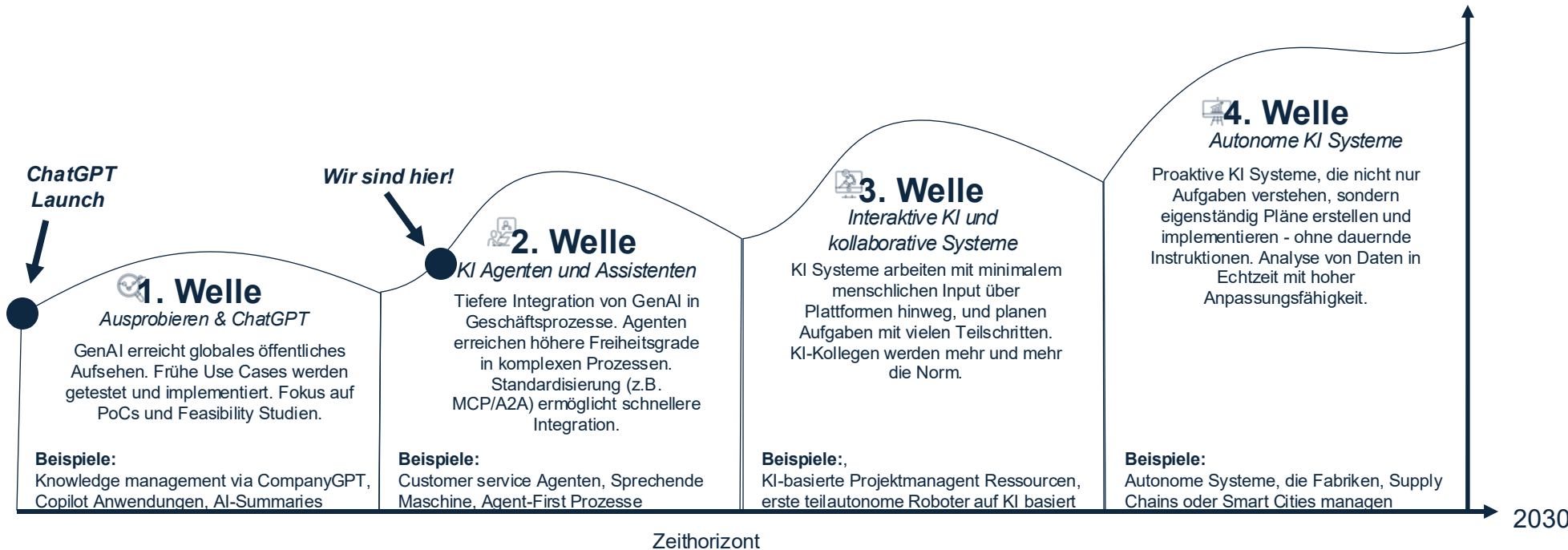


Generative KI

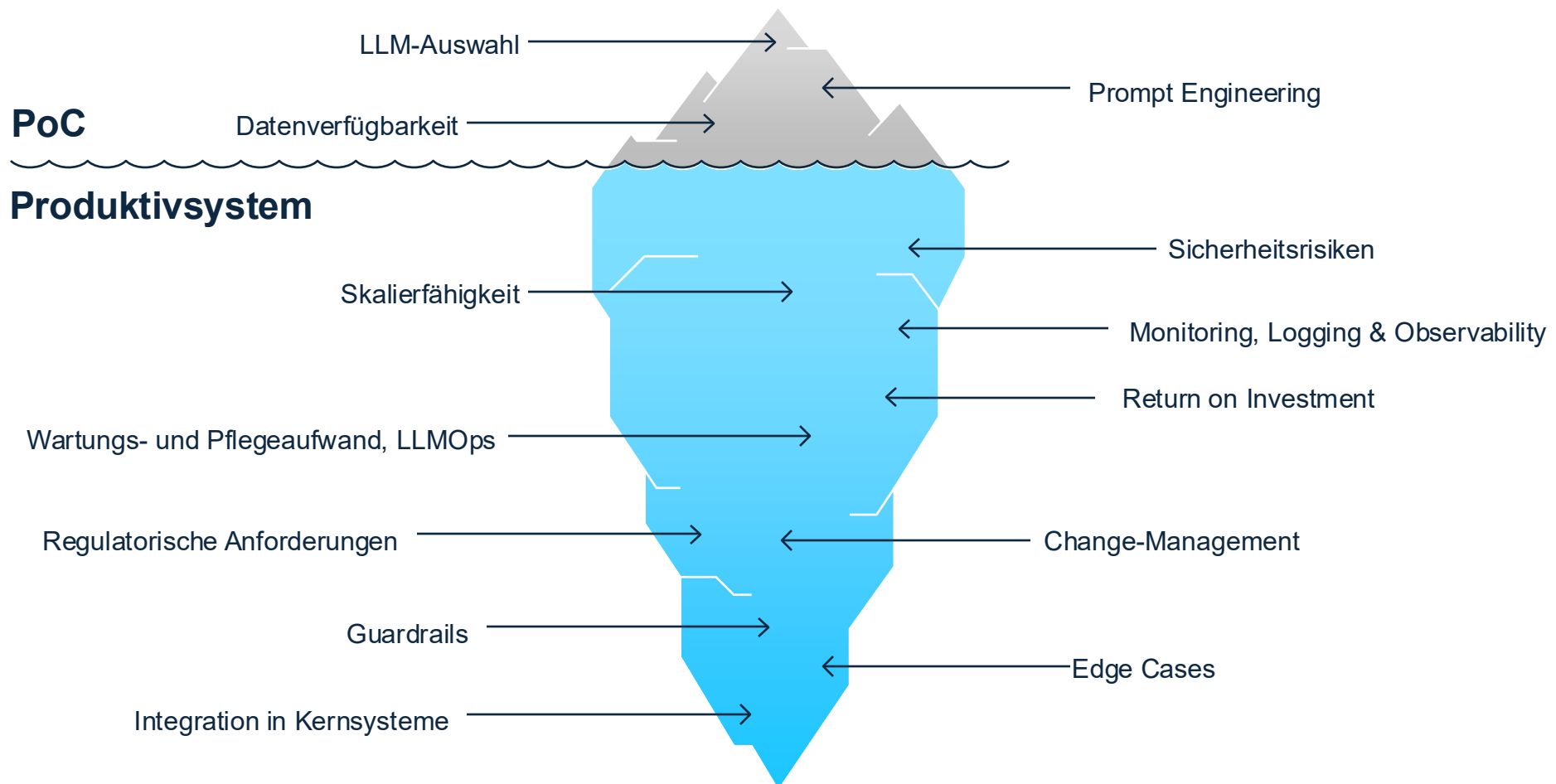
Von der Insellösung zur digitalen KI-gestützten
Gesamtlösung für KMU



Die 4 Wellen der generativen KI



Vom Pilot zur Produktivlösung



Die drei Ebenen eines KI-Stacks

Software

Tools für Orchestrierung, Evaluation, Observability, Routing, Security, Memory, Protokolle, Data Storage etc.



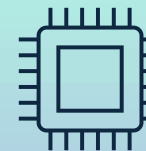
Foundational Models

Leistungsfähige LLMs z.b. Multimodale LLMs, LRMs, SMLs



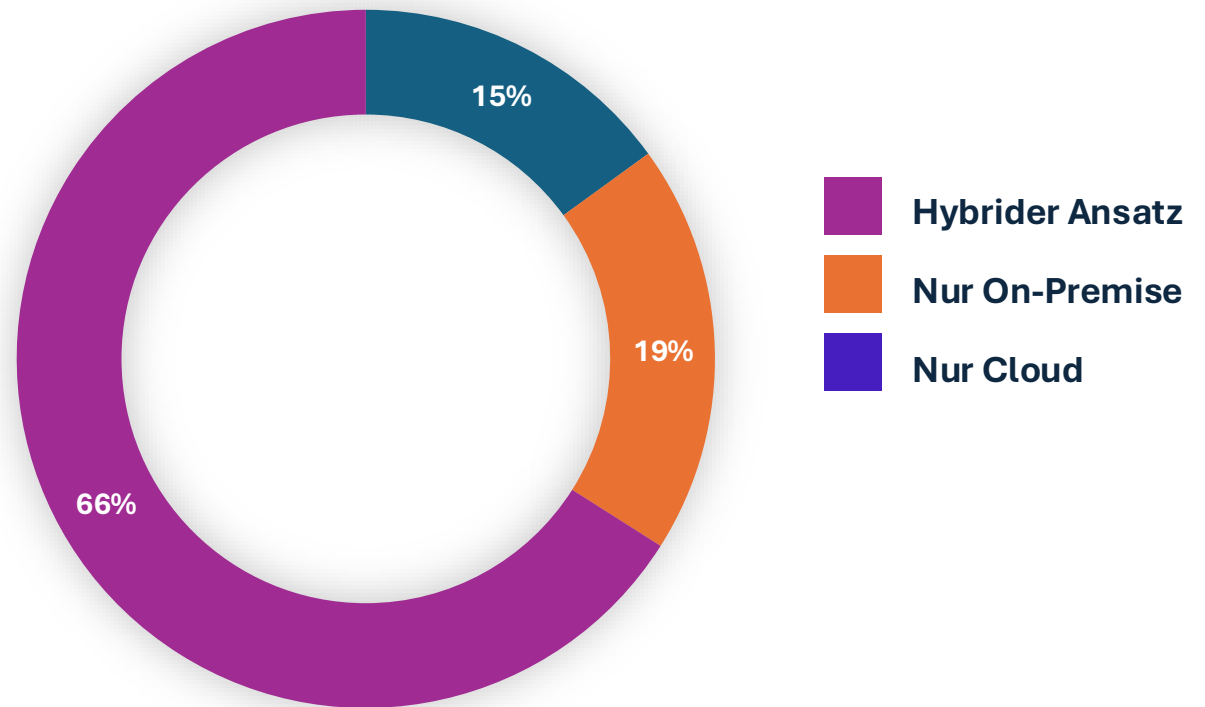
Hardware & Infrastruktur

Leistungsfähige GPUs zum betreiben der Applikationen, On-Prem oder Private / Public Cloud



Hybride Ansätze liegen vorne

„Wie nutzen Sie GenAI
in Ihrem Unternehmen?“



Überblick der Hardware-Angebote

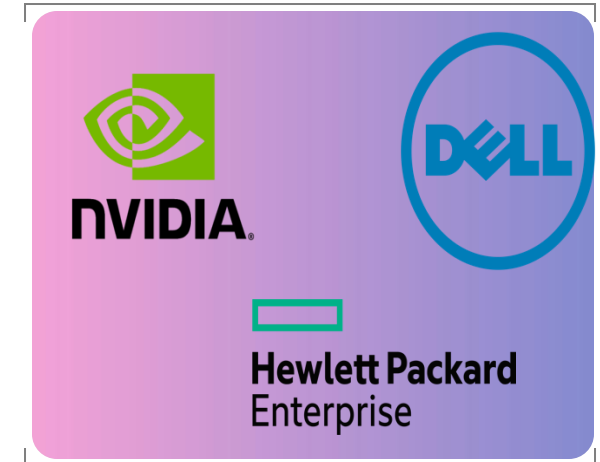
Hyperscaler



Souveräne Cloud



On-Premise



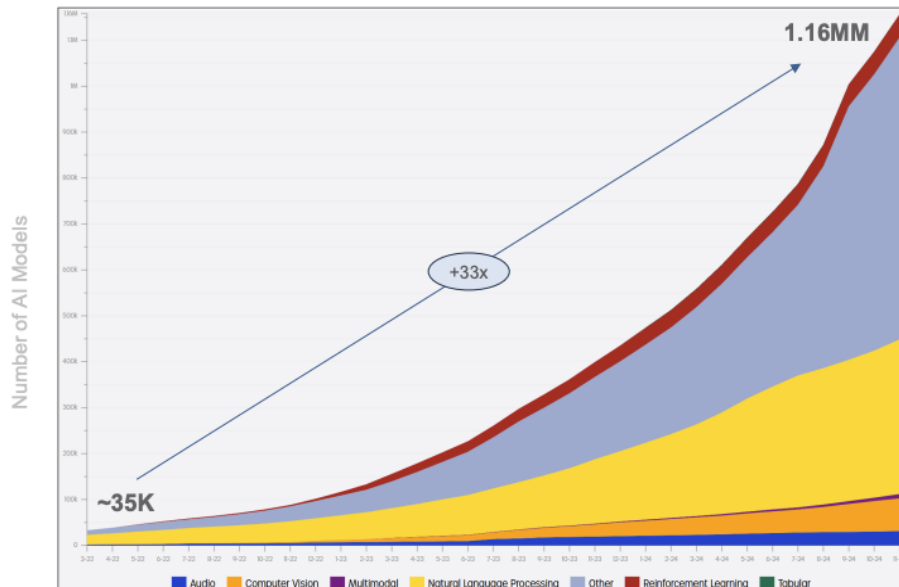
Skalierung

Datensensibilität

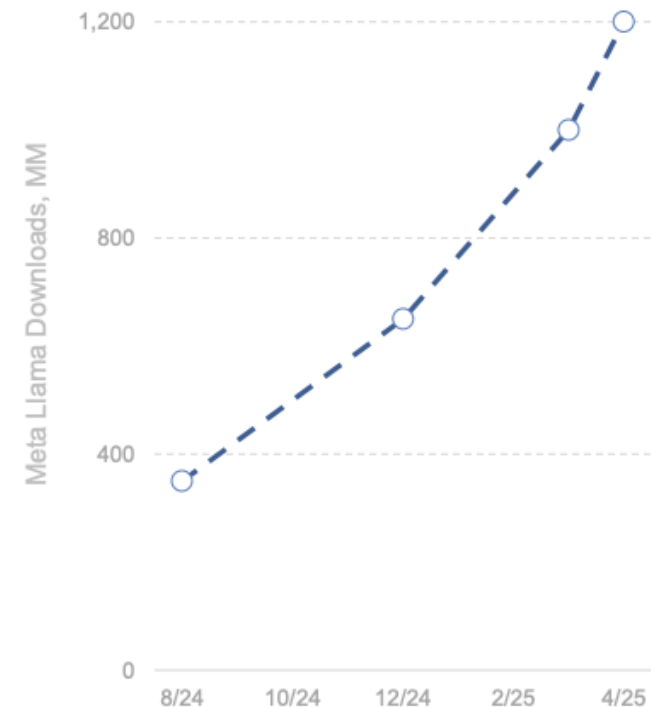


Open-Source Modelle explodieren

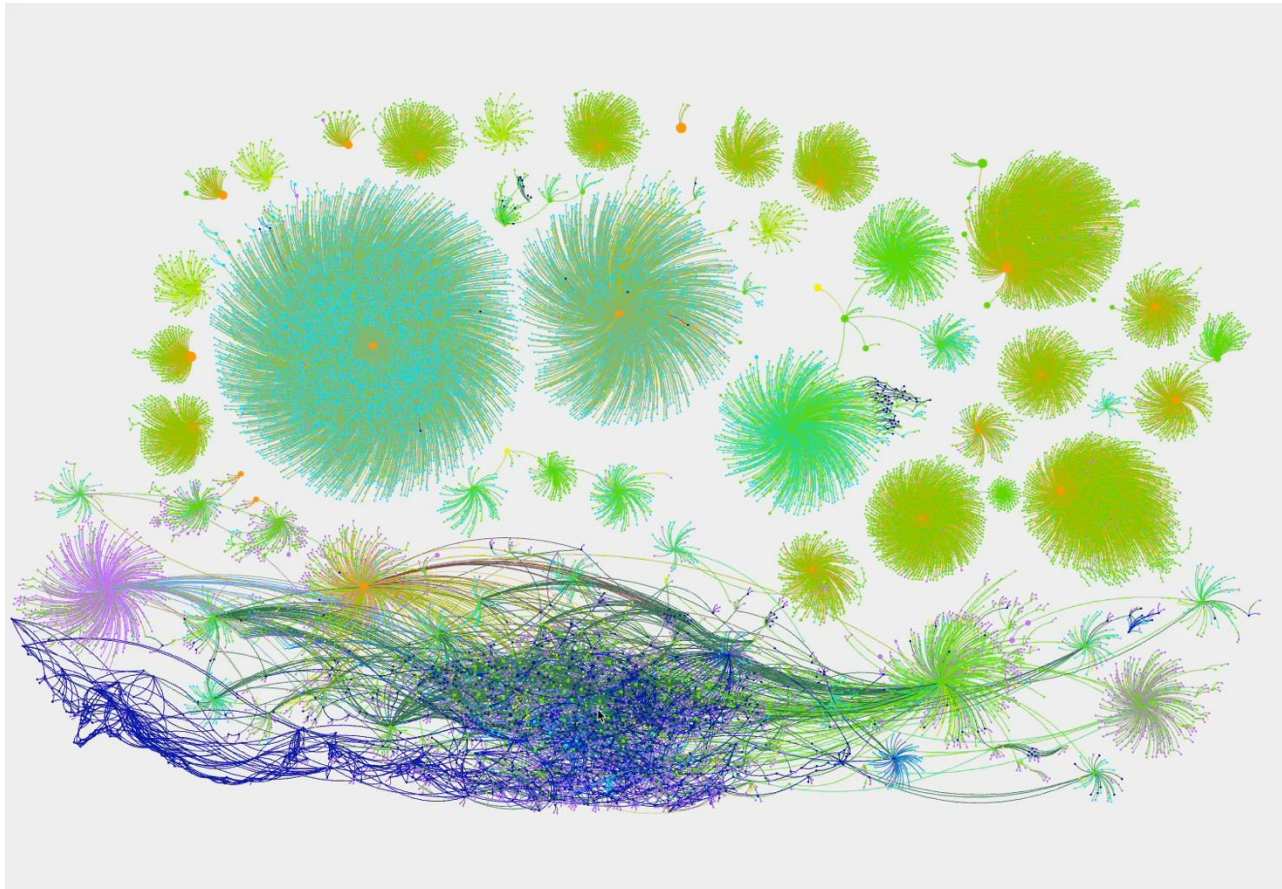
AI Models Available from Hugging Face – 3/22-11/24, per Hugging Face



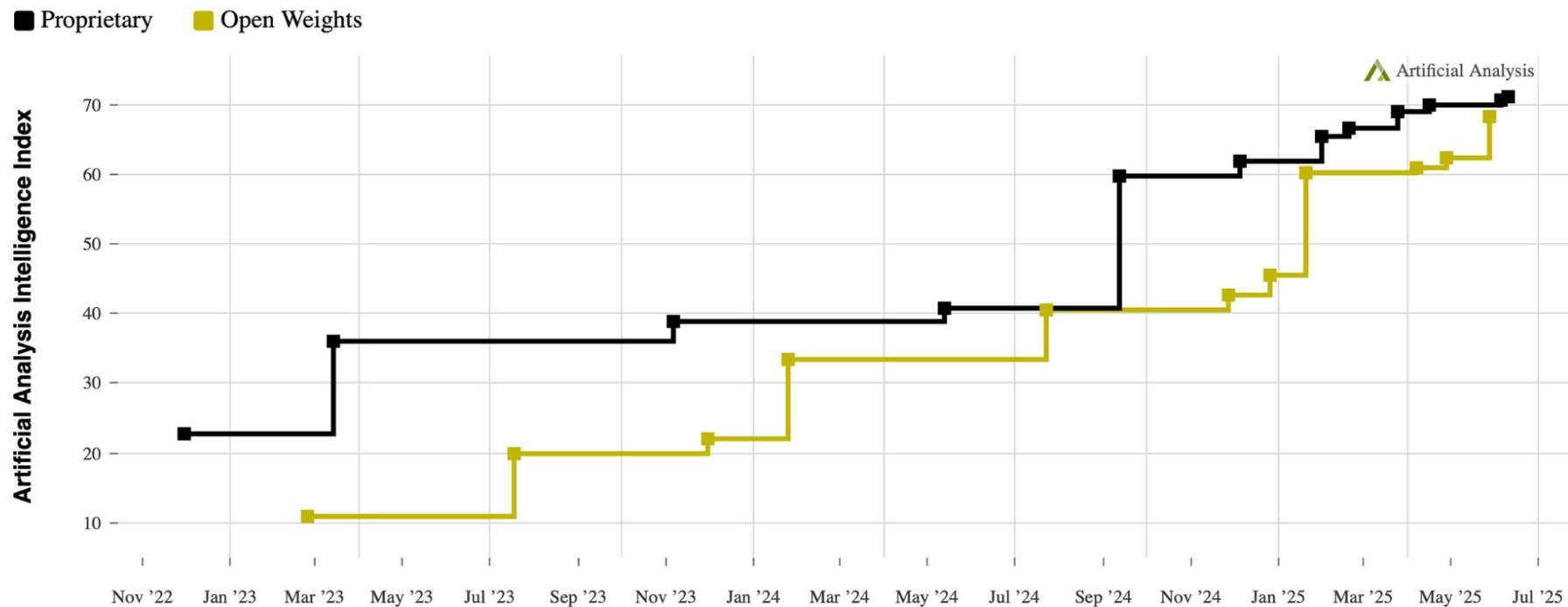
Meta Llama Downloads (MM) – 8/24-4/25



Visualisierung der Modellvielfalt



Open-Source Performance steigt



Softwarevielwelt in allen Ebenen

Data Storage 

Hyperscaler	Closed	Open Source
		
Azure AI Search	Pinecone	ChromaDB

Inferenz Serving 

Hyperscaler	Closed	Open Source
		
Azure OpenAI	Hugging Face Inference Endpoints	Ollama

Orchestration 

Hyperscaler	Closed	Open Source
		
VertexAI Workflows	LangGraph	Haystack

Security & Guardrails 

Hyperscaler	Closed	Open Source
		
AWS Bedrock Guardrails	Lakera	Guardrails AI

Fine Tuning 

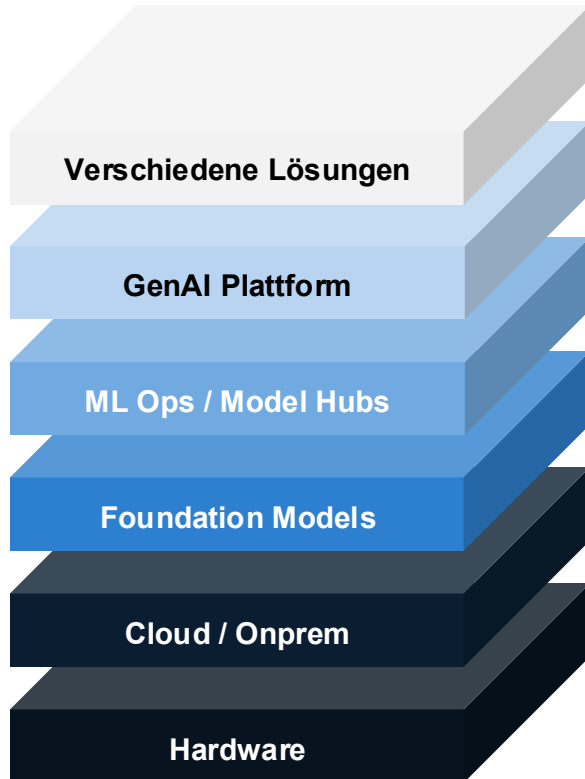
Hyperscaler	Closed	Open Source
		
AzureML	Weights & Biases	Unsloth

Observability & Evaluation 

Hyperscaler	Closed	Open Source
		
VertexAI Monitoring	Langsmith	Langfuse



KI-Plattform als Gesamtlösung



Vorteile einer Plattform

„Ramp Up“ um sechs Wochen reduziert

Der vollständig vor-konfigurierte Docker-container ermöglicht eine geringstmögliche Einrichtungszeit der Plattform, und stellt den schnellstmöglichen Aufbau von Use Cases sicher-standardsiert durch Templates und Best Practices auch bei begrenzter Expertise.

Keine umständliche Datenaufbereitung

Mit den bereits vorhandenen Optionen für plattforminterne Vektordatenbanken, bedarf es keiner langfristigen Vorbereitung für Datenmigrationsschritte.

Kompatibel mit allen LLMs

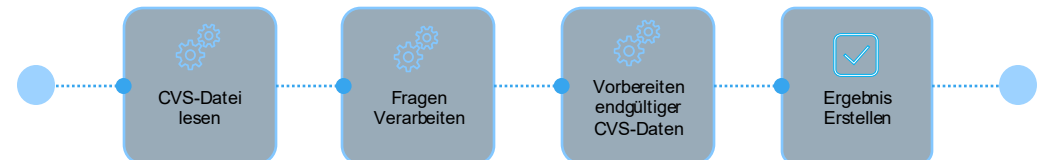
Alle LLMs können via API genutzt werden, wodurch eine vielseitige Kompatibilität und erweiterte Funktionalität über verschiedene KI-Anwendungen hinweg sichergestellt ist

Sicherheit und Kontrolle

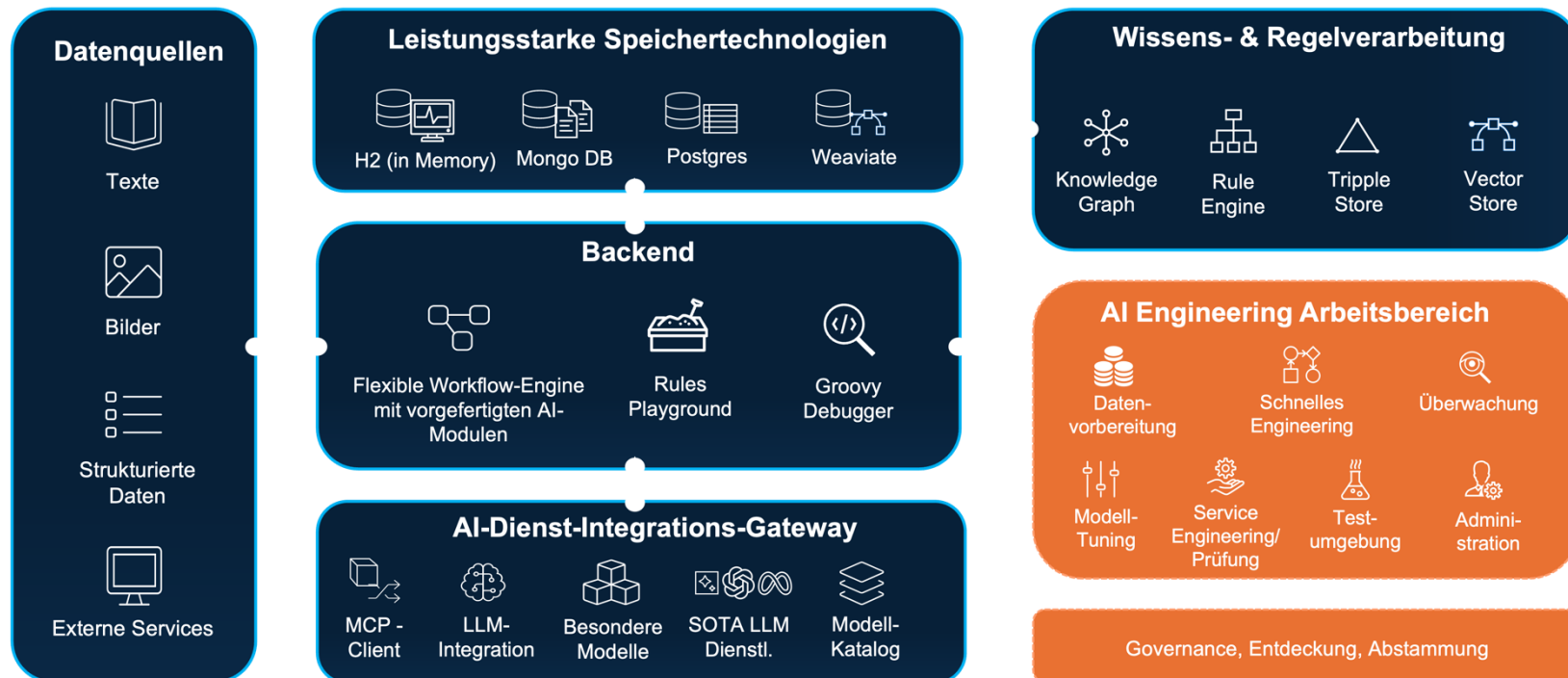
DSGVO-konform, on-prem oder hybrid, mit vollständiger Auditierbarkeit - alles "by design" und kein Extra-Projekt

Low Code / No Code

Modellierung einer KI-Anwendung mit Komponenten über einen hierarchischen Workflow – **beispielhafter Ablauf in der Vertragsanalyse**



Beispielhafte Plattform-Architektur



Alle Beiträge des
IT- & Digitalisierungstages
finden Sie online unter
www.it-tag-38.de
zum Download.

