RheinlandPfalz
MINISTERIUM DES INNERN UND FÜR SPORT

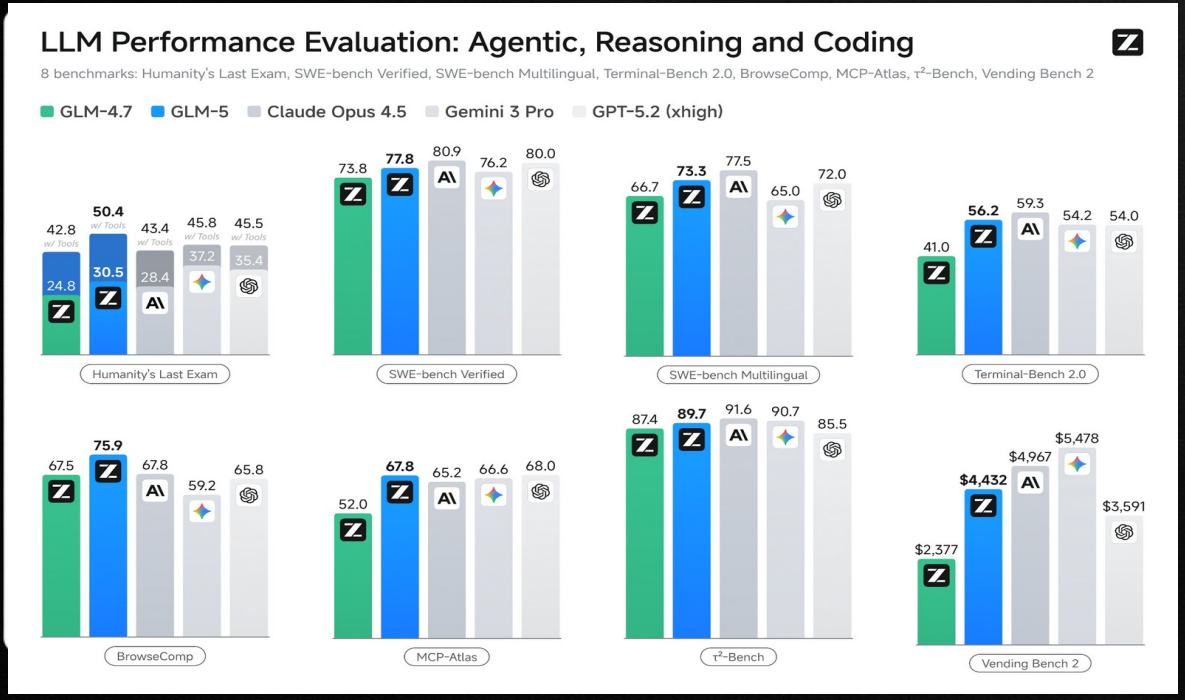
Souveräne KI-Infrastruktur für den Verbandsgemeinde-Verbund

Strategische Implementierung einer DSGVO-konformen On-Premise Lösung auf GPT-5 Niveau.

Finanziert durch die IKZ-Pilotförderung des Landes Rheinland-Pfalz.

Unser Anspruch: State-of-the-Art Intelligenz ohne Cloud-Zwang

Reasoning & Coding Performance & Multi-Task Benchmarks

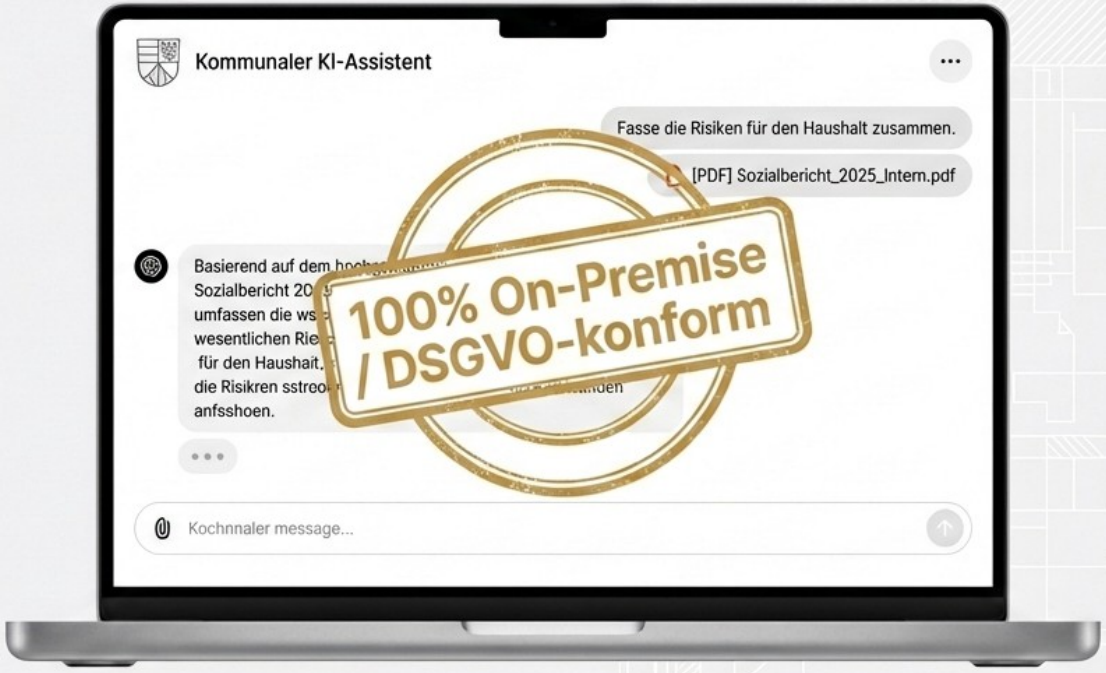


GLM-5 (745 Mrd. Parameter) ist das erste Open-Source-Modell, das mit proprietärel proprietären US-Modellen auf Augenhöhe agiert.

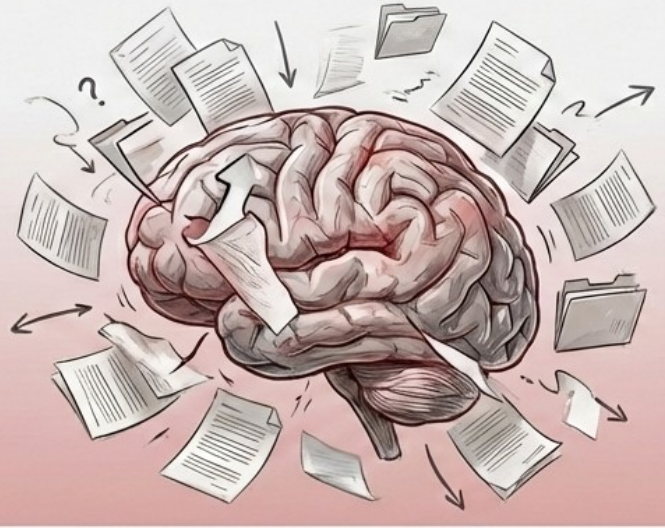
Dank Mixture-of-Experts (MoE) Architektur kombinieren wir Weltwissen mit maximaler Effizienz.

Der Use Case: Ein Omnimodaler Chatbot

- Verarbeitung von sensiblen Akten und Daten.
- Keine Cloud-APIs. Keine Datenabflüsse.
- Ein performanter Assistent für 400 Mitarbeiter.
- Interface, Leistung und Qualität vergleichbar mit GPT-5.



Das Problem: "Context Rot" und die Grenzen des Wissens



Context Rot / Lost in the Middle

Das Problem: ‘Context Rot’ und die Grenzen des Wissens. Auch modernste KI kann nicht tausende Akten gleichzeitig “im Kopf” behalten. Bei zu viel Datenmasse leidet die Präzision – die KI beginnt zu halluzinieren.

Präzision

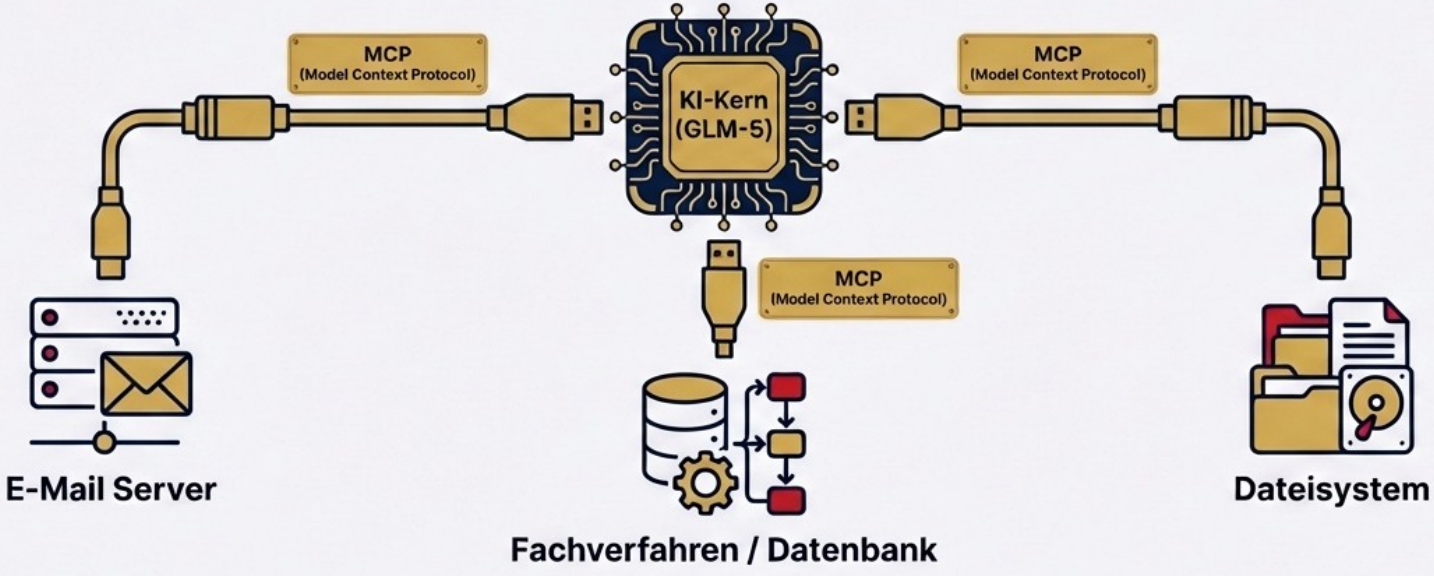
Wir müssen der KI nicht alles geben, sondern nur das Relevante. Vergleichbar mit einem Bibliothekar, der gezielt die passenden Informationen aus einem großen Bestand heraussucht.

Die Lösung: Retrieval-Augmented Generation (RAG)

Eine vorgeschaltete Suchmaschine findet exakt passende Textstellen in unseren Dokumenten. Die KI generiert die Antwort ausschließlich auf Basis dieser **verifizierten Ausschnitte**. Minimierung von **Halluzinationen** garantiert.

Die Integration: Model Context Protocol (MCP)

MCP ist die universelle Schnittstelle und der moderne Ersatz zur API, der es der KI erlaubt, sicher mit unseren Fachverfahren zu kommunizieren. Die KI kann nicht nur lesen, sondern unter Aufsicht Aktionen ausführen.



Der Tech-Stack: Eine Enterprise-Architektur

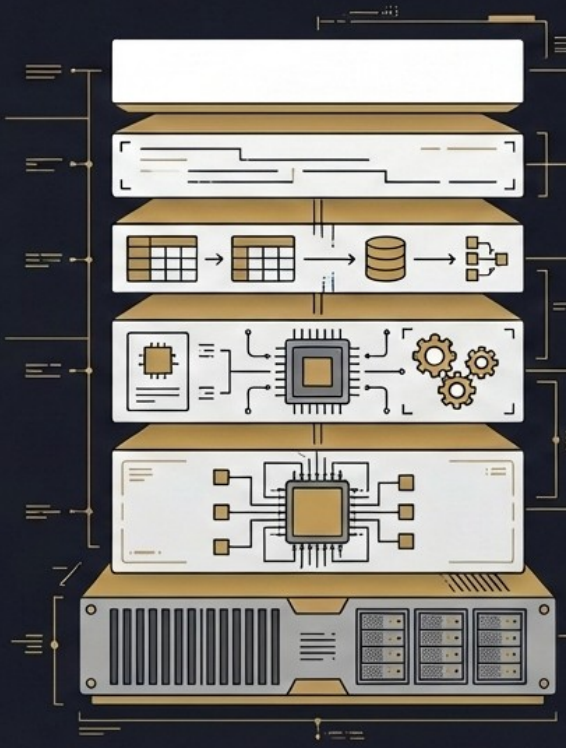
Maximale Performance durch vLLM und FP8-Quantisierung.
Ein vollständig kontrollierbarer Stack ohne Black-Boxes.

Orchestrierung (LangChain)

Steuert den gesamten Ablauf, verbindet die Komponenten und verwaltet den Informationsfluss zwischen Nutzer, Datenbank und KI-Modell.

RAG & Vektordatenbank (Milvus)

Speichert Dokumente als Vektoren für effiziente Ähnlichkeitssuche und liefert relevanten Kontext für die Antwortgenerierung.



Frontend / UI (Chatbot)

Orchestrierung (LangChain)

RAG & Vektordatenbank (Milvus)

Inference Engine (vLLM / SGLang)

Modell (GLM-5 - 745B MoE)

Hardware
(NVIDIA Enterprise Cluster)

Warum Consumer-Hardware scheitert

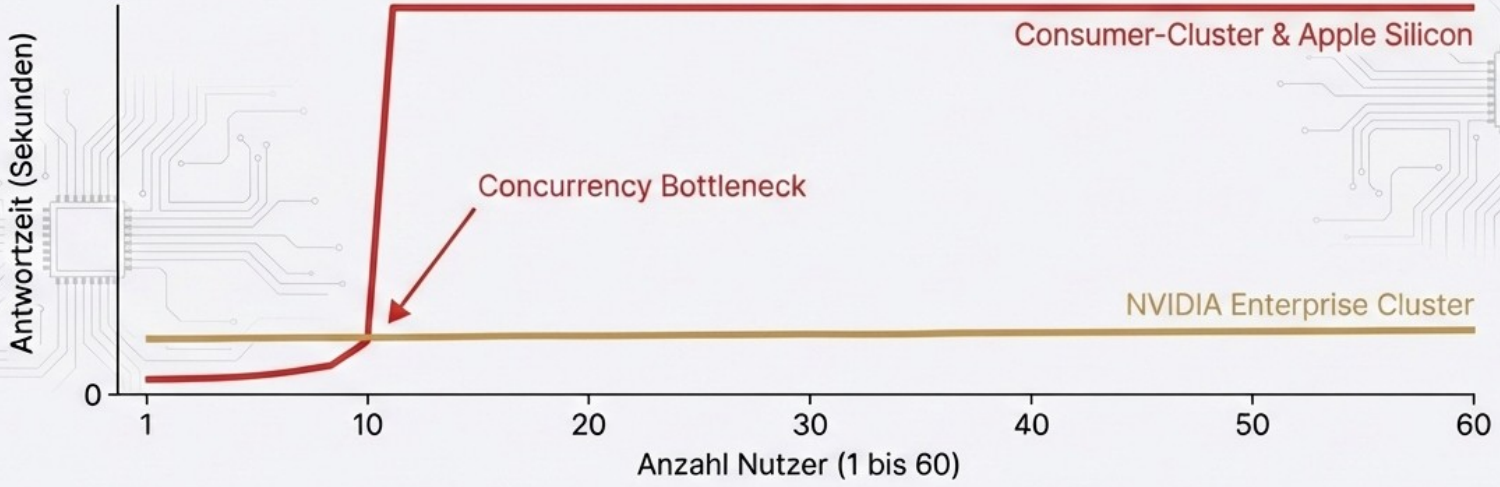
Selbst 24GB pro Karte reichen nicht. Der Flaschenhals ist die Bandbreite.
Wir benötigen Server-Dauerlastfestigkeit und ECC-Fehlerkorrektur.

Hinweis: Frankenstein-Cluster aus Consumer-Hardware haben zwar genügend Speicher im Verbund, aber der Durchsatz und die Komplexität sind nicht geeignet für Concurrency.

Warum Apple Silicon und Consumer-Hardware-Cluster nicht ausreichen

Sowohl Macs als auch Consumer-Cluster brechen bei parallelen Anfragen ein. Es fehlt die Unterstützung für Enterprise-Batching-Technologien wie vLLM und die notwendige Interconnect-Bandbreite.

Antwortzeit bei steigender Nutzerzahl



Die Lösung: H200 HGX & Alternativen

Das Optimum: NVIDIA H200 HGX

- Perfekt für maximale Performance & Skalierbarkeit.
- Ultra-schneller Speicher (HBM3e).
- Preislich jedoch im Premium-Segment (geschätzt > 250.000 € pro System).

Kosteneffiziente Alternativen & Geschätzte Preise

- NVIDIA RTX 6000 Ada Cluster (ca. 60.000 € - 100.000 €)
- L40S Server Cluster (ca. 120.000 € - 180.000 €)
- A100 (80GB) Refurbished Cluster (ca. 80.000 € - 150.000 €)

Die H200 HGX ist die ideale, aber teure Lösung. Die Alternativen bieten einen pragmatischen Kompromiss aus Leistung und Kosten für den stabilen Betrieb des 745 Mrd. Parameter Modells.

Die Finanzierung: 100% Deckung durch IKZ-Förderung

Die Verbandsgemeinden erhalten eine High-End-Infrastruktur **ohne Belastung der eigenen Haushalte.**

Der Wirtschaftlichkeitsnachweis

Vergleich: ChatGPT Pro Kosten für 400 Mitarbeiter über 5 Jahre

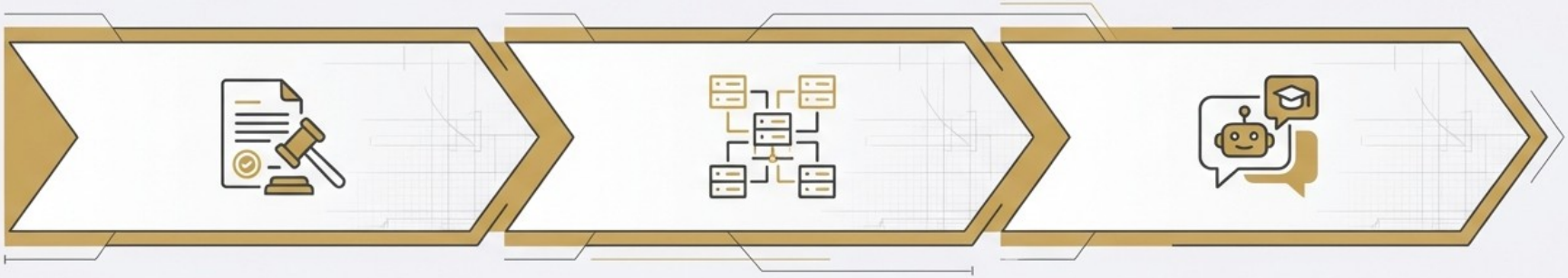
ChatGPT Pro Lizenz: ~20€ / Monat
400 Mitarbeiter x 20€ x 12 Monate = 96.000€ / Jahr
96.000€ / Jahr x 5 Jahre = 480.000€

Wir erreichen den **gleichen** oder höheren Effizienzgewinn **ohne** diese externen **Lizenzkosten**. Unser eigenes System bietet zudem den entscheidenden Mehrwert der Digitalen Souveränität.

0€ Lizenzkosten ✓

Roadmap zur Implementierung

Antragstellung durch eine VG stellvertretend für den Verbund.
Projektlaufzeit mind. 5 Jahre.



Phase 1: Antrag & Beschluss
(Q4 2025)

Phase 2: Beschaffung & Installation
(Hardware Cluster)

Phase 3: Rollout & Schulung
(Pilot-Chatbot)

Ausblick: Vom Chatbot zur Prozessautomatisierung

Die Hardware erlaubt den Betrieb eigener MCP-Tools. Wir bauen eigene 'Helfer', die in jede Fachanwendung gehen und vorhandene Schnittstellen nutzen oder gar selbst erschaffen können.

Auto-Protokollant

Bauantrags-Prüfung

Bürger-Assistenz

Die Transformation der Verwaltung

Weg von der manuellen Texterstellung, hin zur qualifizierten Entscheidung. Schnellere Bürgerdienste und entlastete Mitarbeiter.

Fazit: Wegfall von langwierigen, monotonen Aufgaben ohne Anspruch. Es bleiben nur höherwertige Tätigkeiten übrig, was zu schnelleren Bürgerdiensten und entlasteten Mitarbeitern führt.

Zusammenfassung & Nächste Schritte

Technologie

GLM-5 auf Enterprise Cluster (GPT-5 Niveau).

Finanzen

~370.000 € IKZ-Förderung (100% gedeckt).

Sicherheit

100% On-Premise & DSGVO-konform.

Handlungsbedarf: Ratsbeschlüsse herbeiführen und Antragsfristen beachten.