Projekt R.O.M.

Die souveräne KI-Plattform

ChatGPT-Niveau für die öffentliche Verwaltung. 100 % lokal. 100 % DSGVO-konform.

100 % lokaler Betrieb im eigenen Rechenzentrum

100 % DSGVO-konform durch physische Datenhoheit

Logik und Sprachverständnis auf ChatGPT-Niveau

Das Souveränitäts-Paradoxon lähmt die kommunale Handlungsfähigkeit

Die Privatwirtschaft

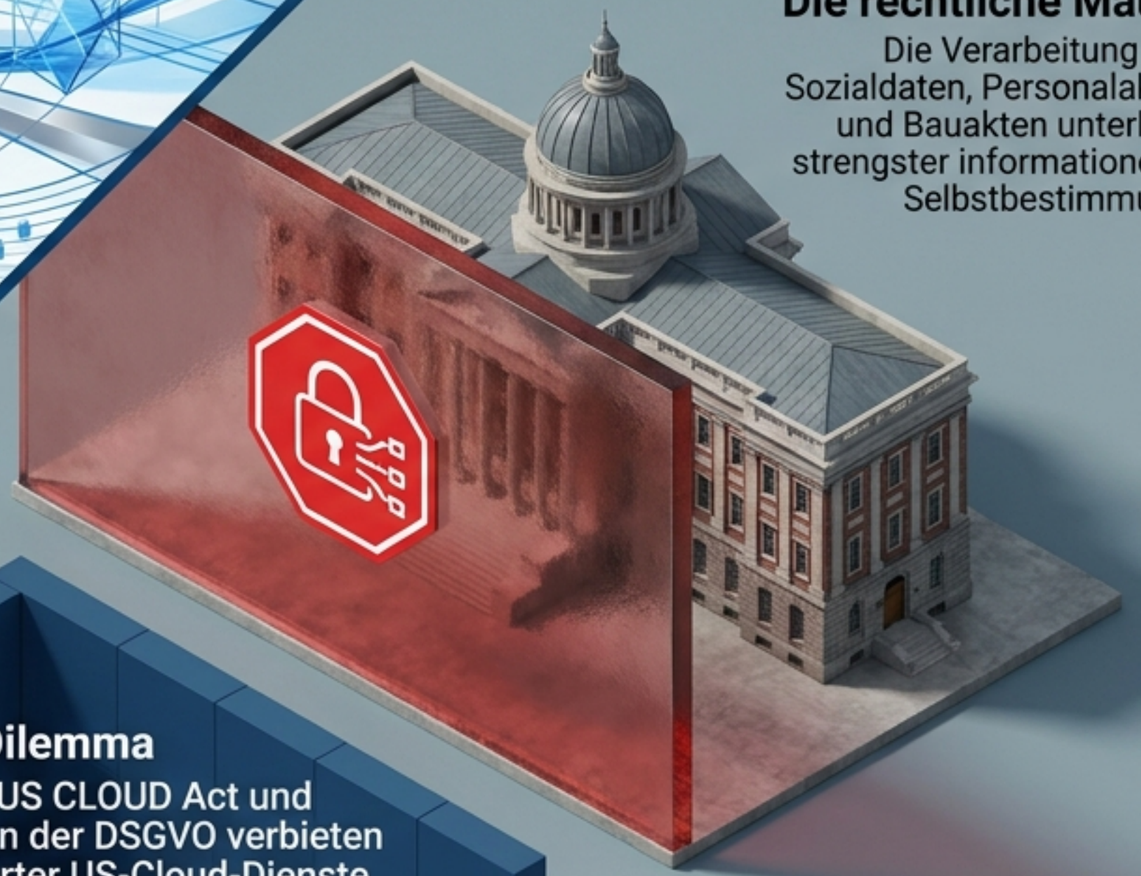
Der demografische Druck

Eine massive Pensionierungswelle trifft auf einen dramatischen Fachkräftemangel. Die Effizienzsteigerung durch generative KI ist keine Option, sondern eine Überlebensfrage für Behörden.

Der öffentliche Sektor

Die rechtliche Mauer

Die Verarbeitung von Sozialdaten, Personalakten und Bauakten unterliegt strengster informationeller Selbstbestimmung.



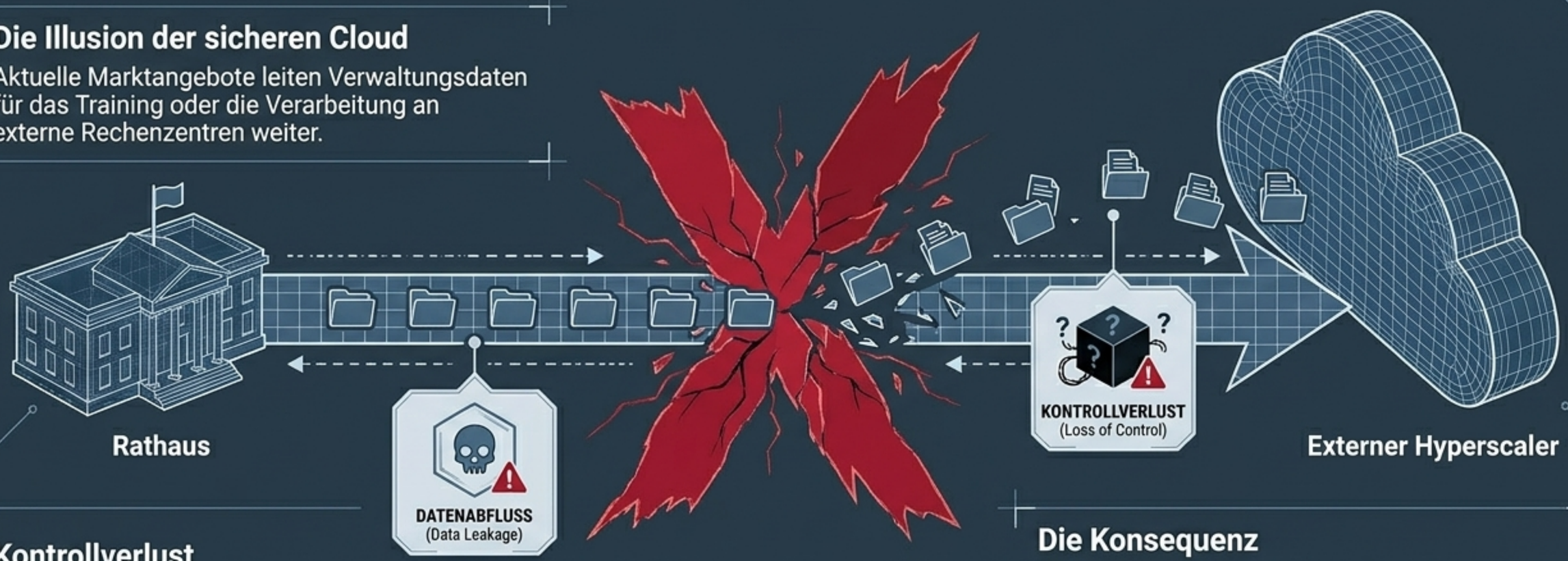
Das Dilemma

Schrems-II, der US CLOUD Act und die strikten Vorgaben der DSGVO verbieten die Nutzung etablierter US-Cloud-Dienste für diese Kernprozesse faktisch vollständig.

Sensible Kernprozesse vertragen keine Kompromisse bei der Datenhoheit

Die Illusion der sicheren Cloud

Aktuelle Marktangebote leiten Verwaltungsdaten für das Training oder die Verarbeitung an externe Rechenzentren weiter.



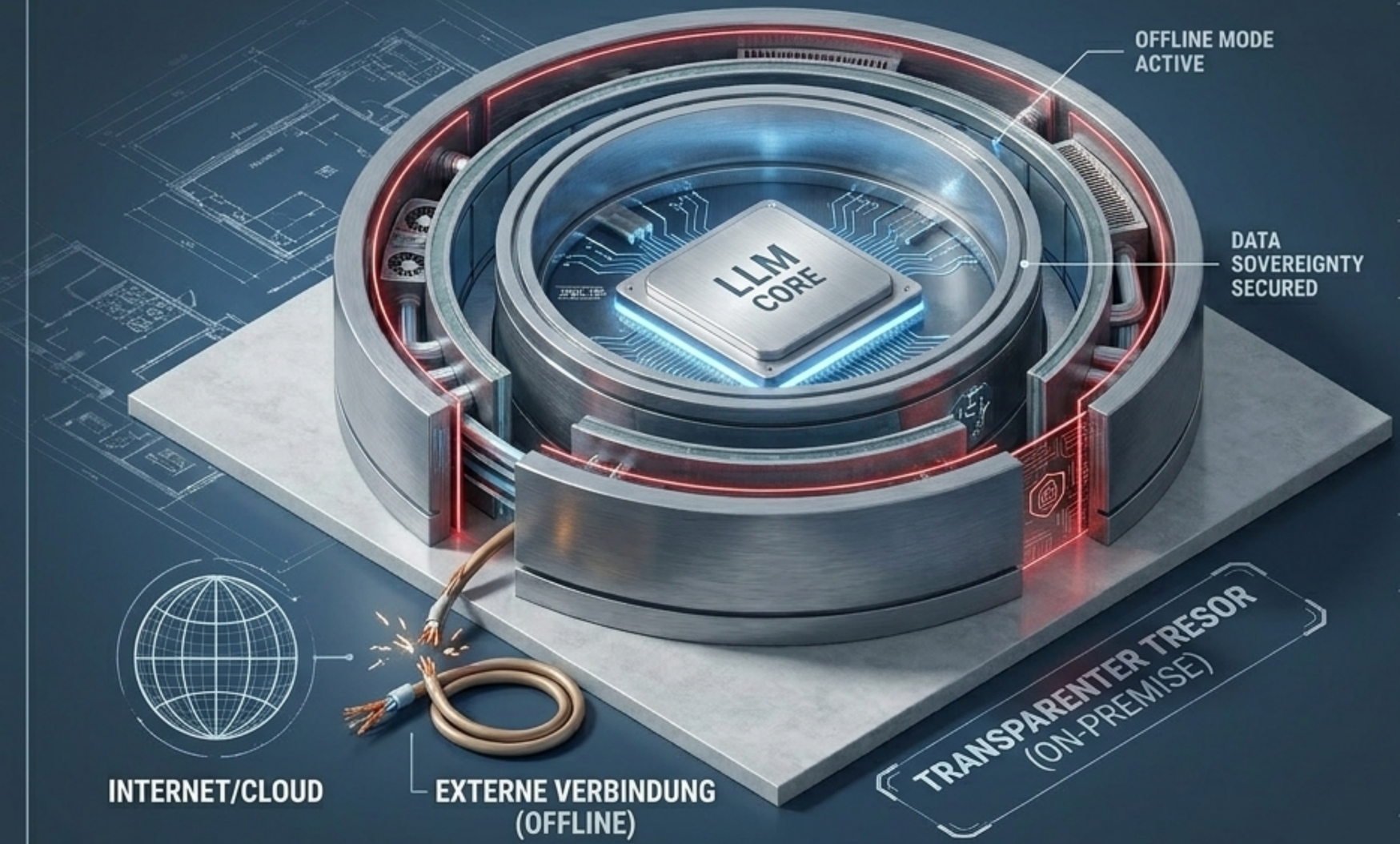
Kontrollverlust

Verlässt ein einziges Datenpaket das Haus, ist die souveräne Kontrolle unwiderruflich verloren (Privacy by Design wird verletzt).

Die Konsequenz

Wer externe Standard-LLMs nutzt, riskiert Datenschutzverstöße bei sensiblen Personendaten oder zwingt die Verwaltung zum technologischen Stillstand (Schatten-IT).

Das Regionale On-Premise-Modell (R.O.M.) bringt die Intelligenz zu den Daten



Der Paradigmenwechsel

Wir senden keine Akten mehr an eine externe KI. Wir bauen ein Frontier-Modell physisch im eigenen, kommunalen Rechenzentrum (SWT Trier) auf.

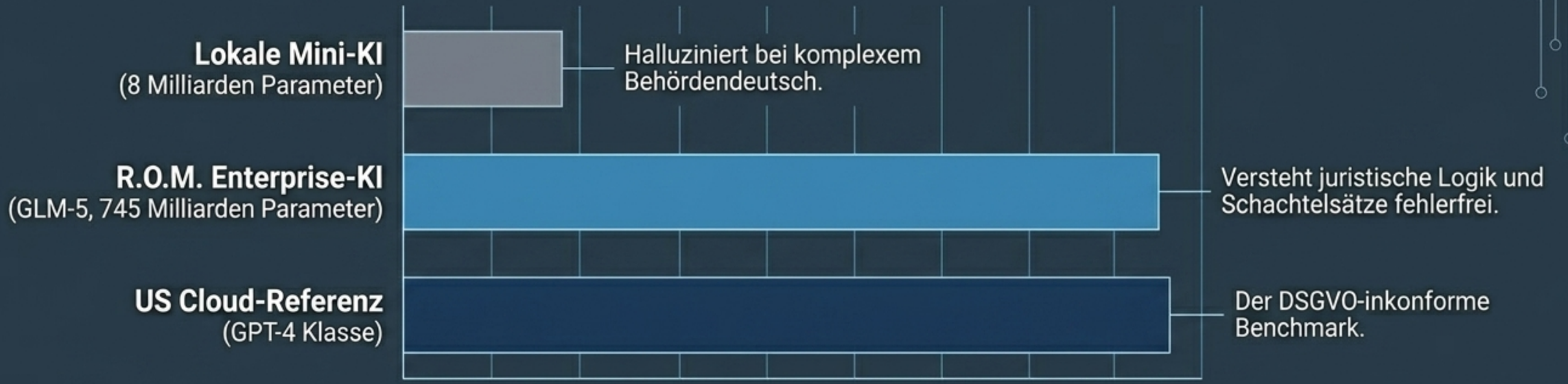
Absolute Isolation

Das System arbeitet zu 100 % offline. Es gibt keine externe Cloud-Anbindung, keine Telemetrie und keinen Datenabfluss.

Souveräne Basis

R.O.M. ist kein IT-Experiment, sondern die zwingend notwendige Basisinfrastruktur für die datenschutzsichere Handlungsfähigkeit der Verwaltung.

Nutzerakzeptanz erfordert State-of-the-Art Intelligenz ohne Kompromisse



Die Lücke schließen

Open-Weight-Modelle der Spitzenklasse (Frontier-Modelle) haben massiv aufgeholt.

Warum 745 Milliarden Parameter?

Eine lokal betriebene KI wird nur genutzt, wenn sie so gut ist wie die privaten Tools der Mitarbeiter.

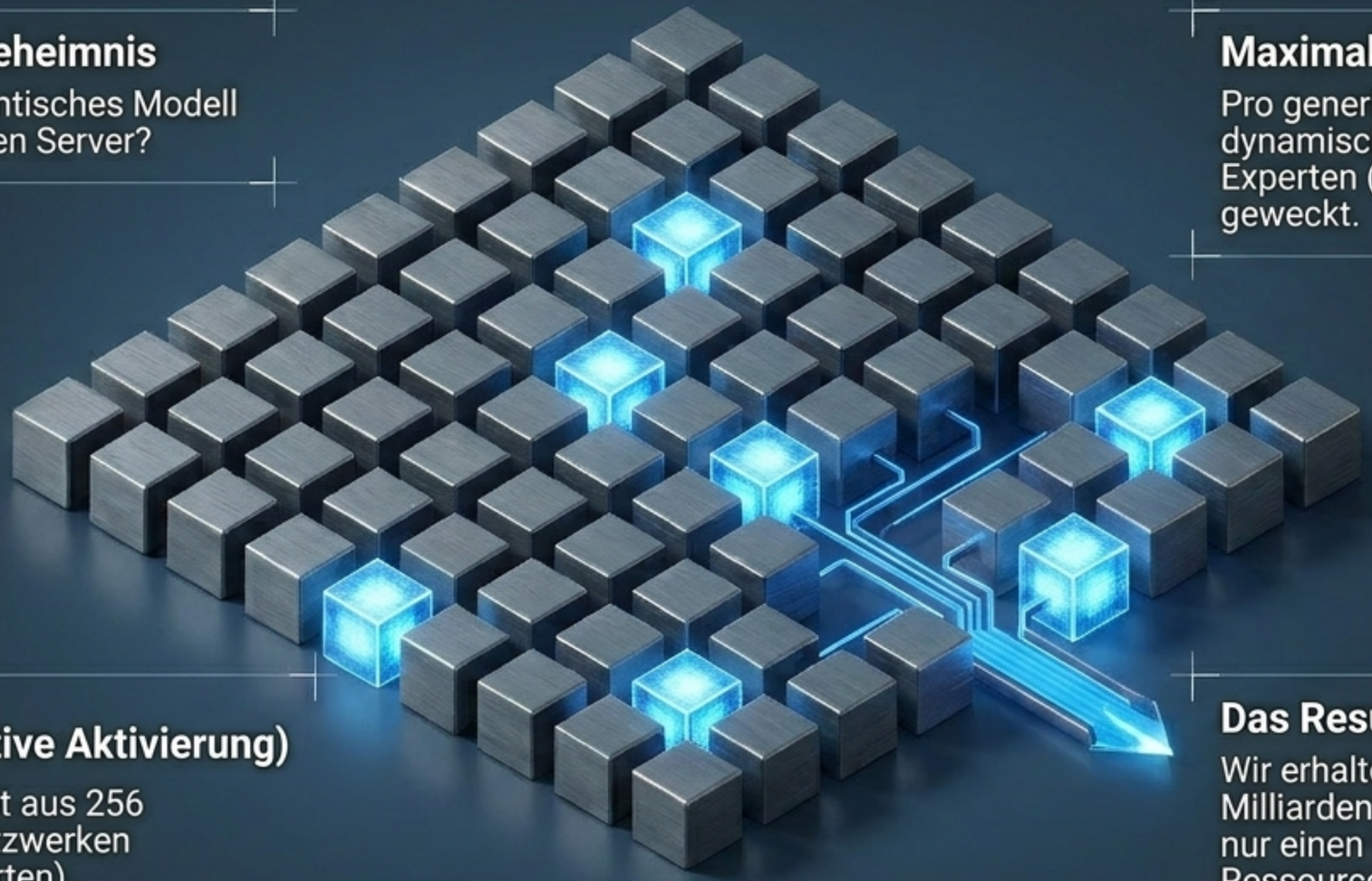
Kognitive Qualität

Nur Modelle dieser Größenordnung durchdringen widersprüchliche Gesetzeslagen und komplexe Bauakten, ohne Fakten zu erfinden.

Hochleistung durch Spezialisierung: Die Mixture of Experts Architektur

Das Effizienz-Geheimnis

Wie passt ein gigantisches Modell (GLM-5) auf unseren Server?



Maximale Rechenleistung

Pro generiertem Wort werden dynamisch nur die 8 relevantesten Experten (ca. 40 Mrd. Parameter) geweckt.

Sparsity (Selektive Aktivierung)

Das Modell besteht aus 256 spezialisierten Netzwerken Netzwerken (Experten).

Das Resultat

Wir erhalten das Wissen aus 745 Milliarden Parametern, benötigen aber nur einen Bruchteil der Hardware-Ressourcen für die Berechnung.

Aktenwissen sicher anbinden durch Retrieval-Augmented Generation (RAG)

Kein riskantes Nachtraining

Das Sprachmodell lernt keine sensiblen Akten auswendig (verhindert Datenlecks im Modellgedächtnis).

Das Spickzettel-Prinzip

Das System sucht in Millisekunden die passenden Akten aus der Vektordatenbank und reicht sie der KI in einem 8.000-Token-Kontextfenster an.

Garantierte Belegbarkeit

Die KI generiert Wissen nicht aus dem Nichts, sondern stützt sich ausschließlich auf die verifizierten internen Dokumente.

Kompromisslose Mandantenfähigkeit durch Database-Level Isolation

4 Kommunen, 1 Server, 0 Datenvermischung

R.O.M. bedient die beteiligten Verbandsgemeinden auf derselben Hardware.

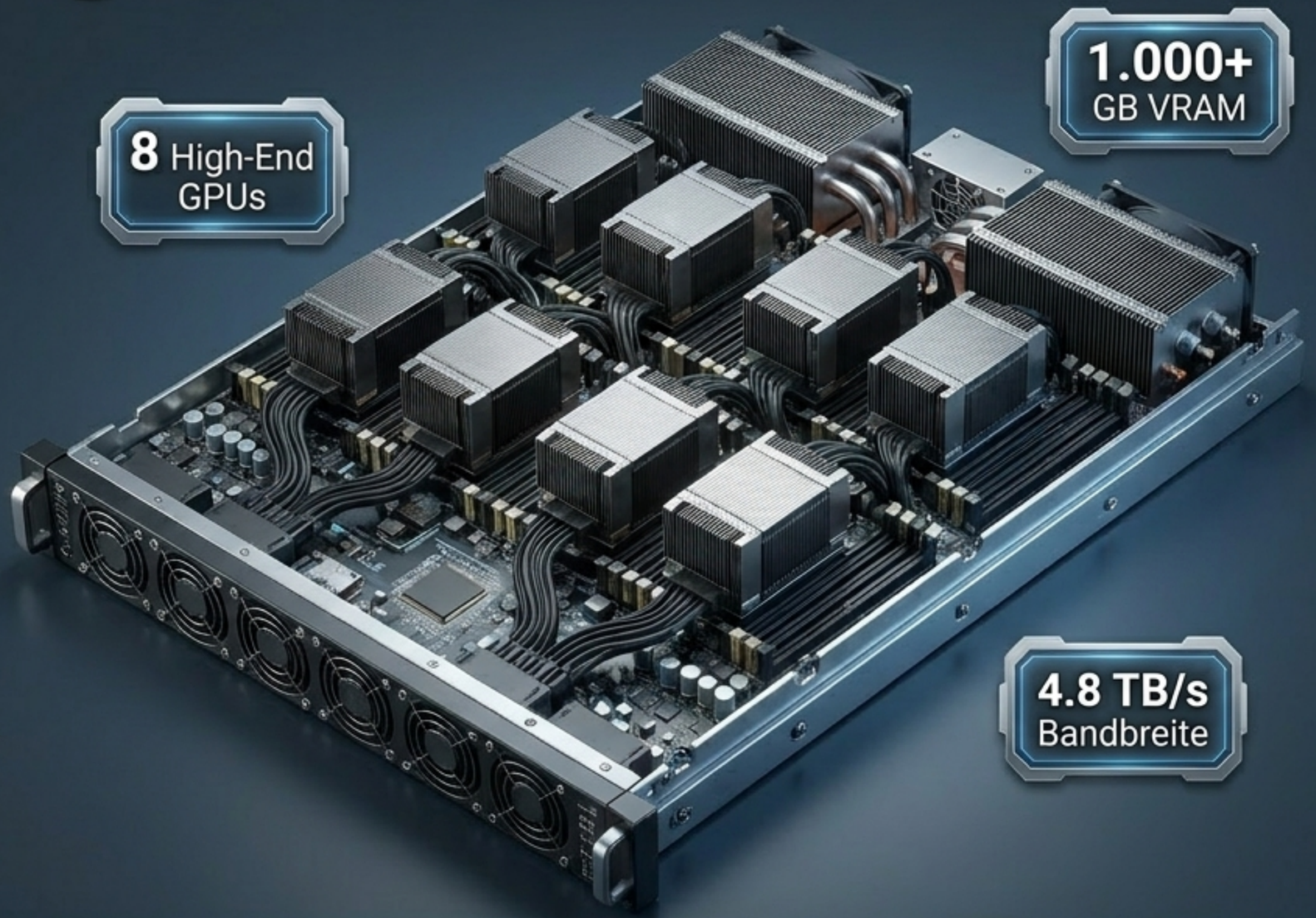
Tiefe Isolierung

Um DSGVO-konform zu bleiben, wird auf der tiefsten Datenbankebene getrennt (Database-Level Isolation).

Need-to-Know-Prinzip

Das Berechtigungskonzept des bestehenden Fileservers (Active Directory) gilt 1:1. Die KI durchsucht nur Dokumente, für die der jeweilige Nutzer explizite Leserechte besitzt.

Der physische Maschinenraum der digitalen Souveränität



Warum Enterprise-Hardware?

Um 60 Mitarbeiter gleichzeitig ohne Warteschlangen flüssig bedienen zu können (15–30 Token pro Sekunde), ist extreme VRAM-Kapazität zwingend erforderlich.

Keine Bastellösung

Konsumenten-PCs scheitern an der Netzwerk-Kommunikation bei Modellen dieser Größe.

Der Investitionsbedarf

Ein Enterprise-Inferenzserver bildet das ausfallsichere Fundament. Kostenschätzung: ca. 300.000 – 350.000 Euro. Eine Investition, die das Projekt vor dem Scheitern an Ressourcen-Engpässen bewahrt.

Architektur-Vergleich: Warum R.O.M. alternativlos ist

	US-Cloud-KI (ChatGPT)	Lokale Mini-KI (Büro-PC)	R.O.M. Enterprise On-Premise
100 % DSGVO-Konformität bei Sozialdaten	✗	✓	✓
Kognitive Logik auf ChatGPT-Niveau	✓	✗	✓
Volle Kontrolle über Vektordaten	✗	✓	✓
Kein Vendor Lock-In (Abo-Fallen)	✗	✓	✓

Cloud-Lösungen fallen bei der Datenhoheit durch. Lokale Bastellösungen fallen beim Logik-Niveau durch. Nur R.O.M. erfüllt das Souveränitäts-Paradoxon.

Finanzierbarkeit durch Interkommunale Zusammenarbeit (IKZ)

Der Sweet-Spot der Auslastung

Ein dedizierter Inferenzserver ist für eine Kommune allein überdimensioniert. Im Verbund von vier Kommunen erreichen wir maximale Wirtschaftlichkeit.

Kostenteilung

Gebündeltes IT-Know-how und geteilte Betriebskosten (Strom, Wartung) entlasten die kommunalen Haushalte.

Maximale Förderkulisse

Das Land Rheinland-Pfalz fördert solche IKZ-Infrastrukturprojekte zur digitalen Souveränität mit bis zu 100 %.

Entwicklungs-Roadmap: Von der Basis zur Automation

Stufe 1: Fundament

Phase 1: Souveräner ChatGPT-Ersatz. Infrastrukturaufbau. Die sichere Plattform zur internen Texterstellung, Übersetzung und Strukturierung.

Stufe 2: Wissen

Phase 2: Das Verwaltungsgehirn (RAG). Anbindung der internen Dokumentenmanagementsysteme. Der Chatbot liest, versteht und zitiert lokale Akten.

Stufe 3: Automation

Phase 3: Agentische Systeme. Nutzung des Model Context Protocols (MCP) zur Automatisierung mehrstufiger Verwaltungsprozesse und Etablierung externer Bürger-Schnittstellen.

Souveränität ist ein System, kein Software-Update

Rechtliche Governance
(DSGVO & IKZ-Mandanten)

State-of-the-Art Software
(vLLM & GLM-5)

Massive Hardware
(1TB VRAM)

Wer R.O.M. nutzt, kauft kein isoliertes Tool.

Er investiert in den dauerhaften Erhalt der hoheitlichen Handlungsfähigkeit.

Hardware, Software und Governance greifen wie Zahnräder ineinander.

Fehlt eine Schicht, bricht das System der Datensicherheit zusammen.

Fazit & Beschlussvorschlag: Wir übernehmen die Kontrolle

Die Entscheidung

Wir warten nicht auf **unsichere Cloud-Lösungen,** sondern bauen unsere **eigene, zukunftssichere Basis-Infrastruktur** im Rahmen des **Deutschland Tech Stacks**.

Nächste Schritte:

1. **Grundsatzbeschluss zur interkommunalen Zusammenarbeit (IKZ) fassen.**
2. **Die VG Schweich als federführende Kommune mandatieren.**
3. **Förderantrag beim Land Rheinland-Pfalz einreichen.**