

# KI-Systeme auf eigener Infrastruktur

mit **Open Source** Lösungen am Gesundheitsministerium

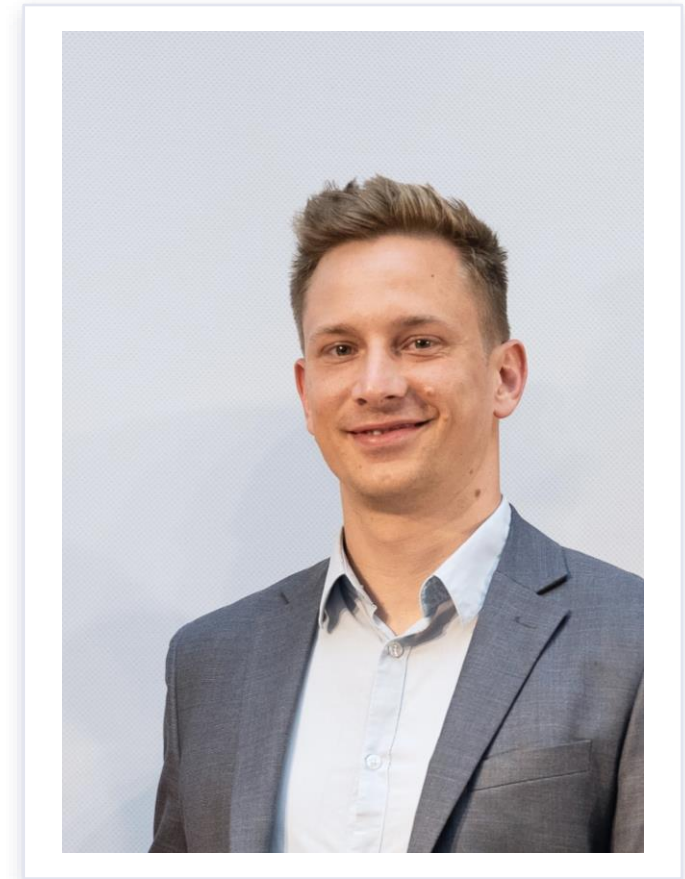
◆ Philipp Schardax, MSc. M.A. · [onpremises.org](https://onpremises.org)

Vortrag bei der OSSBIG — Open Source Software in Business & Government



# Philipp Schardax

- Betreue den Bereich KI im Gesundheitswesen für das BMASGPK (Gesundheitsministerium)
- **Systemadministrator für interne On-Premise KI-Anwendungen**
- Akademischer Hintergrund: Wirtschaftsinformatik & Artificial Intelligence
- Mitglied der Austrian Society for Artificial Intelligence (ASAI)
- **Betreibe die gemeinnützige Informationsplattform onpremises.org** über Open Source Lösungen für KI auf eigener Infrastruktur



# Agenda



**01**

## Ausgangslage

Wofür wollten wir KI nutzen?

**02**

## Digitale Souveränität

Wieso unabhängig auf eigener Infrastruktur?

**03**

## Retrieval Augmented Generation

Die Basis moderner KI-Systeme

**04**

## Umsetzung mit Open Source

zB mit Open WebUI + vLLM

**05**

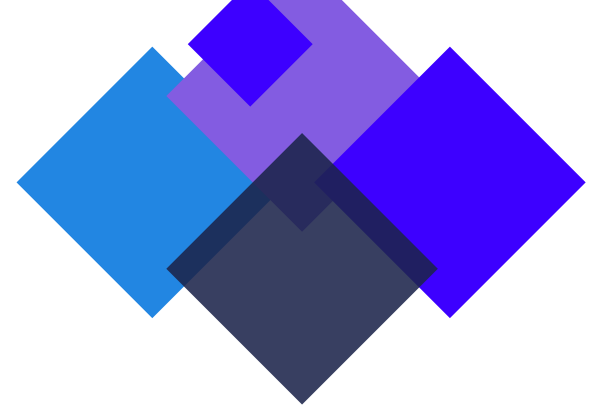
## Erfahrungen am Gesundheitsministerium

Setup, Hardware, Learnings

**06**

## Vorteile & Ausblick

Was Open Source möglich macht



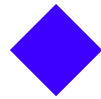
# 01

TEIL 1

## Ausgangslage

Was wollten wir erreichen?

# Klassische Use Cases



## Wissensmanagement

Fragen beantworten auf Basis interner  
Dokumente



## Texte & E-Mails

Antworten und Entwürfe schnell erstellen



## Software Development

Coding-Unterstützung direkt im Editor



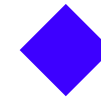
## Prozessautomatisierung

Integration in bestehende Anwendungen



## Transkription

Besprechungen automatisch  
verschriftlichen



## Übersetzungen

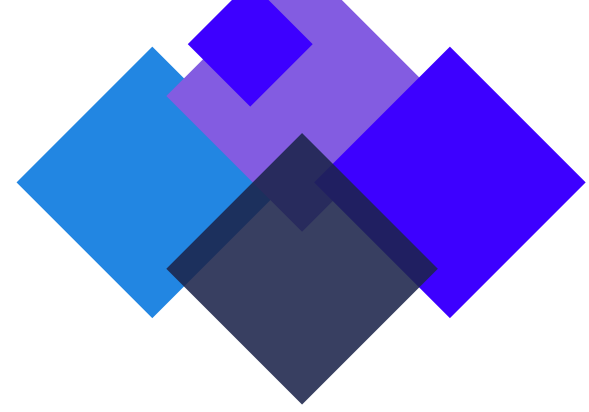
Ganze Dateien on premise übersetzen



# Mitarbeiter nutzen bereits KI!

- In ALLEN Organisationen
- Mitarbeiter nutzen deren **private Zugänge** zu ChatGPT, Gemini, Perplexity etc.
- **Rechtliche Schwierigkeiten, Daten fließen ab, Verlust von Geschäftsgeheimnissen**

**Um das zu verhindern  
braucht es  
saubere Alternativen.**



02

TEIL 2

# Digitale Souveränität

Wieso KI auf eigener Infrastruktur?



# Wieso digital souverän?

- **Unvereinbarkeit von DSGVO und US CLOUD Act**
  - Rechtliche Problematik bei Cloud-Lösungen aller US-Anbieter: potenzieller Zugriff durch US-Behörden
  - Besteht unabhängig vom physischen Standort der Server
- **NEU seit 3.6.2026: Verordnungsvorschlag für den Cloud and AI Development Act (CADA):  
explizites Verbot von Clouds unter Kontrolle von nicht-EU Unternehmen für viele Datenkategorien**
- **Vermeidung von Abhängigkeit (Vendor Lock-in)**
- **Sperren und Exportbeschränkungen:** aktuelles Beispiel Claude Fable 5
- **Bedenkenlose Nutzung aller internen Daten**
  - Daten bleiben auf eigener Infrastruktur — keine Datenschutz-Bedenken

# Cloud and AI Development Act (CADA)

- Vorgestellt am 3.6.2026
- Differenzierte Sicherheitsstufen und Auflagen für Anbieter
- Für Stufen 3 & 4 explizit nur noch EU-Cloud-Anbieter zulässig
- Für Stufen 1 & 2 weiterhin Unvereinbarkeit mit US CLOUD Act
- Zusätzlich: neue Open Source Strategie, geprüfte Repository Sammlung, und Digital Commons EDIC
- **einfache Lösung: on prem bleiben**

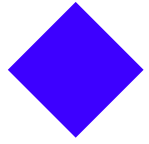


## Welche Daten dürfen in welche Cloud?

Die 4 Sicherheitsstufen (Union Assurance Levels) im Entwurf des Cloud and AI Development Act (CADA) der EU-Kommission vom 3. Juni 2026

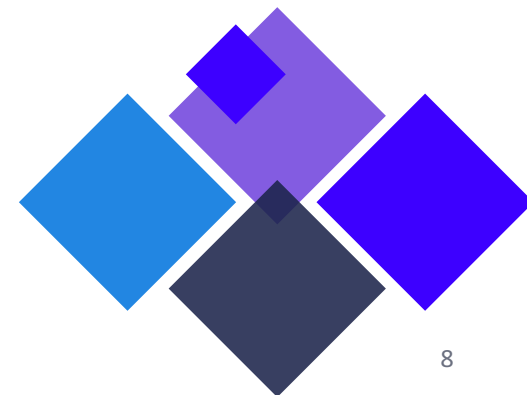


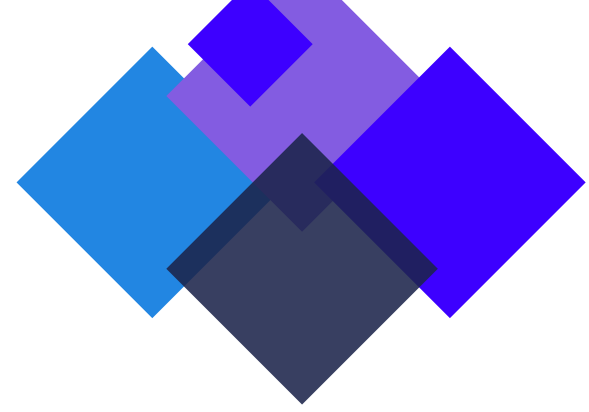
SENSIBILITÄT DER DATEN UND STRENGE DER ANFORDERUNGEN



# Was tun Sie, wenn Ihr Cloud-Anbieter nächstes Jahr die Preise verdreifacht?

Digitale Souveränität ist **kein Goldplating**.





03

TEIL 3

# Retrieval Augmented Generation

Die Basis moderner KI-Systeme



# RAG ist der Schlüssel

- Um Halluzinationen zu vermeiden
- Um auf aktuelles Wissen zuzugreifen
- Um internes Wissen zu nutzen — ohne weiteres Training
- Um KI-Systeme auf leistbarer Infrastruktur zu betreiben

**ChatGPT, Gemini,  
Perplexity, Grok ...**

**nutzen alle RAG.**



# Was RAG macht

Übergibt dem LLM in Echtzeit relevante Informationen, um Prompts zu bearbeiten — und macht neues Wissen zugänglich, ohne neu zu trainieren:



**Interne Dokumente**



**Internetsuche**



**SQL-Datenbanken**

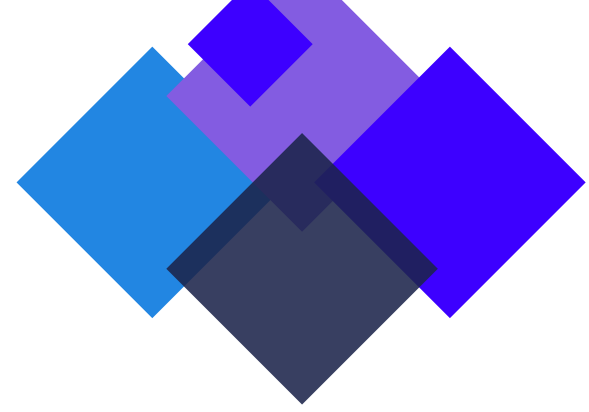


**Bestehende Codebase**



# Die RAG-Architektur





04

TEIL 4

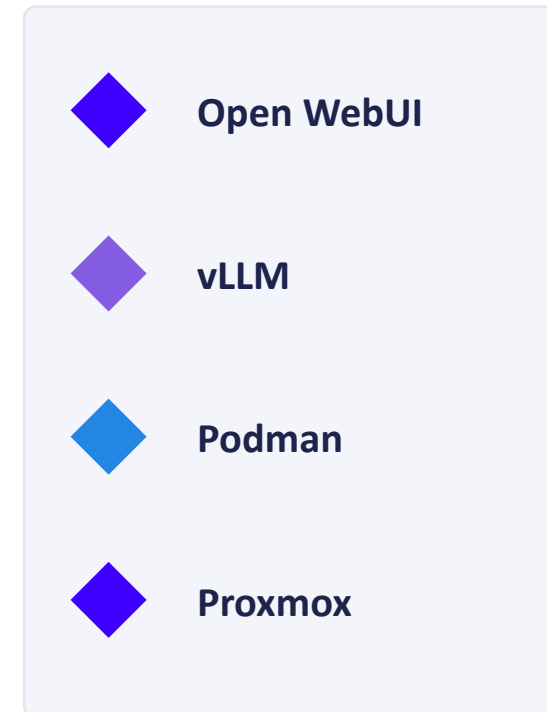
# Umsetzung am Gesundheitsministerium

Open WebUI + vLLM



# Kernstück: Open WebUI + vLLM

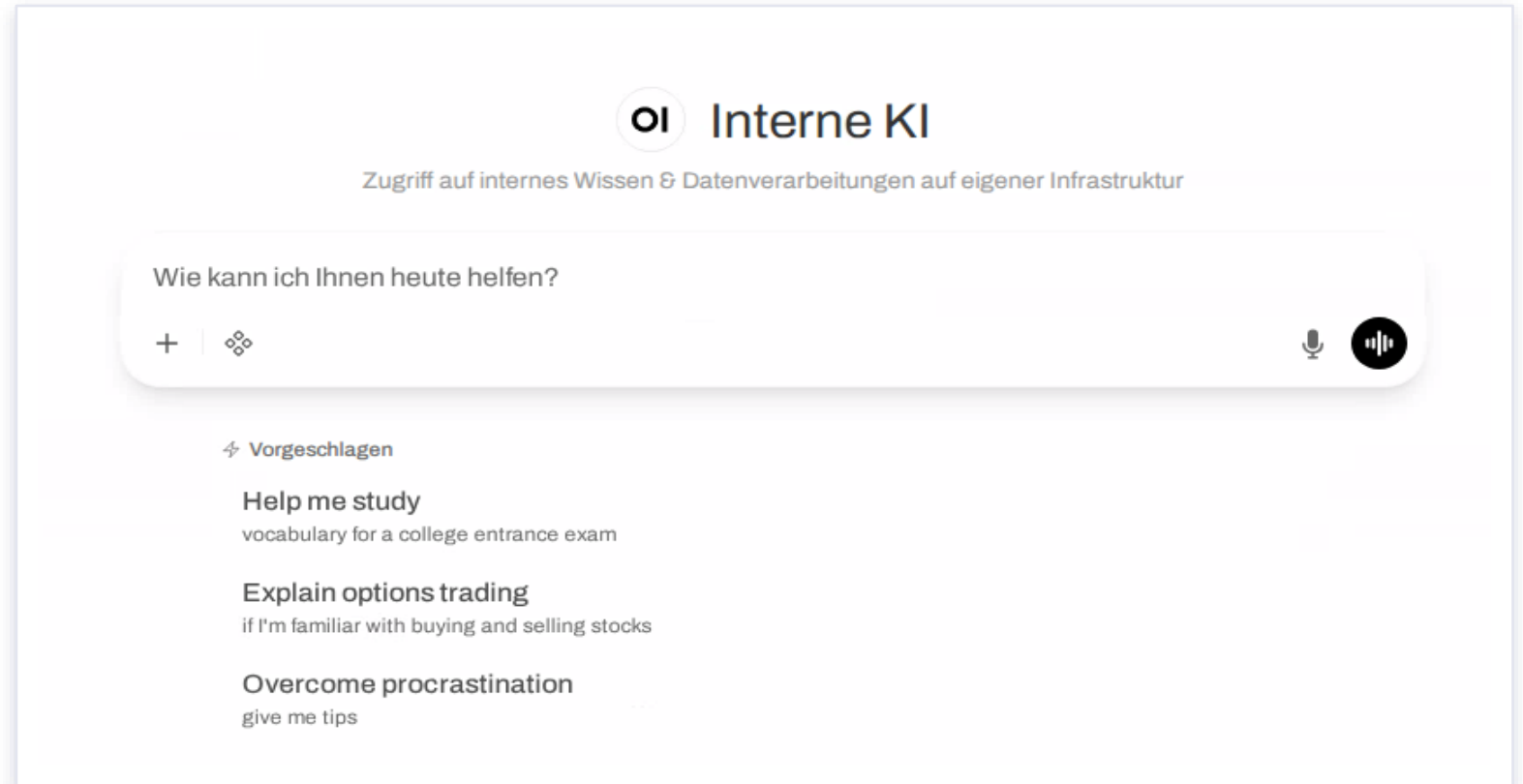
- **Zugriff über den Browser — GUI sehr ähnlich zu ChatGPT**
- Internes Wissen kann per RAG in Masse genutzt werden
- Internet-Suche integriert; Dokumente direkt im Chat hinzufügar
- **Open Source LLMs: GPT-OSS 120B, Mistral, Devstral, Gemma...**
- Läuft vollständig auf eigener Infrastruktur (vLLM als Engine)
  
- Infrastruktur: 4 Stk. RTX 6000 Pro mit je 96 GB VRAM
- **Virtualisierung mit Proxmox → Ubuntu VMs → Podman Container**
- Single Sign On via OIDC Stammportal





# KI auf eigener Infrastruktur — im gewohnten Design

- **Vertraute Chat-Oberfläche**
- Keine Umgewöhnung für Nutzer:innen
- 100 % auf eigener Infrastruktur





# Beispiel: Briefing-Dokumente erstellen

KI X/B GPT-OSS:120b ▾ +

Erstelle ein Briefing Dokument mit ca. 1000 Wörtern über den europäischen Gesundheitsdatenraum

OI KI X/B GPT-OSS:120b  
Nachgedacht für 3 Sekunden ▾  
**Briefing Dokument – European Health Data Space (EHDS)**  
ca. 1 000 Wörter

### 1. Einleitung und strategischer Kontext

Der **European Health Data Space (EHDS)** ist die erste sektorspezifische Datenraum-Initiative der EU und wurde am 3. Mai 2022 von der Europäischen Kommission als Verordnungsvorschlag vorgestellt [2023\\_10\\_06\\_Reede...nvent.docx](#). Er ist Teil der im Februar 2020 veröffentlichten **EU-Datenstrategie**, die den Aufbau gemeinsamer europäischer Datenräume für verschiedene Sektoren vorsieht [2024\\_03\\_13\\_Onep...ealth.docx](#). Der EHDS soll sowohl den **primären** als die **sekundären** Nutzen von Gesundheitsdaten regeln und damit die digitale Gesundheitsversorgung, Forschung und Innovation in Europa nachhaltig stärken.

### 2. Rechtliche Grundlagen

Der Vorschlag basiert auf der **Verordnung (EU) 2022/2481** (HealthData@EU) und ist eng verknüpft mit dem **Data Governance Act (DGA)** [One\\_Pager\\_EHDS...dierf.docx](#). Der Rechtsrahmen definiert die **Pflichten von Datenverantwortlichen**, die Rechte von Bürger\*innen sowie die Governance-Strukturen. Die Verordnung schafft einheitliche Regeln für den grenzüberschreitenden Austausch von Gesundheitsdaten und die **sekundäre Nutzung** zu Forschungs- und Innovationszwecken [2024\\_03\\_13\\_Onep...ealth.docx](#).

### 3. Ziele und Nutzen

ZIEL	BESCHREIBUNG
<b>Bürger*innen-Zugang</b>	Erleichterter, sicherer Zugriff auf eigene Gesundheitsdaten in jedem EU-Staat (z. B. über MyHealth@EU) <a href="#">FAQ Bürger_inne...165258.pdf</a>
<b>Grenzüberschreitende Versorgung</b>	Unterstützung von Telemedizin, EU-Rezepten und e-Prescriptions, sodass Patient*innen im EU-Ausland behandelt werden können <a href="#">FAQ Bürger_inne...165258.pdf</a>
<b>Forschung &amp; Innovation</b>	Strukturierter, rechtssicherer Zugang zu <a href="#">.ßen</a> , standardisierten Datensätzen für sekundäre Nutzung <a href="#">2024_03_13_Onep...ealth.docx</a>

Eine Nachricht senden

+ | 🔗

🎤 🔊

Zugriff auf interne Wissensquellen +  
hochgeladene Kurzbeschreibung.

Performance: 130 token/s

vgl. ChatGPT: ca. 70 token/s



# Delegierbares Wissensmanagement

- Eigene Knowledge Spaces je Abteilung — Fachbereiche pflegen ihr Wissen selbst; die KI greift im Hintergrund darauf zu.

The screenshot displays a web interface for knowledge management. At the top, there are navigation tabs: Modelle, Wissen, Prompts, and Werkzeuge. Below this, the main heading is 'Wissen 22' with a '+ Neues Wissen' button on the right. A search bar contains the text 'Wissen durchsuchen' and a dropdown menu is set to 'Alle'. The main content area shows a grid of knowledge collections, each with a green 'SAMMLUNG' label, a title, an update status, and the creator's name.

Sammlung	Update Status	Ersteller	Sammlung	Update Status	Ersteller
X/A/6	Aktualisiert vor 38 Minuten	Von Philipp Schardax	X/B/10 EU	Aktualisiert vor 5 Tagen	Von Philipp Schardax
European Health Data Space	Aktualisiert vor 5 Tagen	Von Philipp Schardax	MFF	Aktualisiert vor 5 Tagen	Von Philipp Schardax
Tests	Aktualisiert vor 6 Tagen	Von Philipp Schardax	Vorlagen Requirements Engineering	Aktualisiert vor 6 Tagen	Von Philipp Schardax



# Beispiel: Fragebeantwortung mit Quellenangabe

- Fragen direkt aus internem Wissen beantwortet
- Wichtig z. B. bei Urlaubsvertretung
- Mit Quellenangabe und Relevanz-Score

KI X/B GPT-OSS:120b ▾ +

## 7. Ausblick und Bedeutung für Österreich

Für Österreich bedeutet der EHDS eine **Vertiefung der europäischen Gesundheitsdigitalisierung**. Die als nationale Basis-Infrastruktur in das EEHRxS-Schema eingebunden, sodass österreichische Patient:innen MyHealth@EU mit anderen EU-Ländern ausgetauscht werden können [FAQ Bürger\\_inne...165258.pdf](#).

Zudem werden österreichische Forschungseinrichtungen von einem leichteren Zugang zu grenzüberschreitenden standardisierten Gesundheitsdaten profitieren, was die Entwicklung neuer Therapien und digitaler Gesundheitsdienste beschleunigt. Die gesetzliche Verpflichtung, ein **Opt-In** für das EU-Rezept zu implementieren, stärkt die Gesundheit und unterstützt die Vision einer truly European Health Union [FAQ Bürger\\_inne...165258.pdf +1](#).

## 8. Fazit

[240806-eHealth-bf \(1\).pdf](#)

**Inhalt 81.52%** (Seite 23)  
 Potenzial genannt ( OECD/European Observatory on Health Systems Policies 2023 ). (OECD/European Observatory on Health Systems Policies 2023 ). Auf der anderen Seite zeigen Ergebnisse des DESI-Index für das Jahr 2023 im Hinblick auf die elektronische Patientenakte, dass Österreich beim Indikator "Zugang zu Gesundheitsdaten in electronic health records" im EU27-Vergleich über dem Durchschnitt liegt (fünfter Rang) (European Commission 2023a ). Deshalb wird Österreich im Bericht "Digital decade e-realm of health information. In: J Med Internet Res 14/1:e19

**Inhalt 80.47%** (Seite 66)  
 ODERKIRK , Jillian (2021): Survey results: National health data infrastructure and governance. In: OECD Health Working Papers No 127:  
 OECD (2022): Health Data Governance for the Digital Age: Implementing the OECD Recommendation on Health Data Governance. Hg. v. Publishing, OECD. Paris

8 2023\_10\_06\_Redebeitrag\_IKT\_Konvent.docx  
 9 2024\_03\_13\_Onepager\_HBM\_Bilaterales\_Arbeitsgesprach\_Australien\_digital\_health.docx  
 10 SC\_Hausreither\_Rede\_GDHP\_eHealth\_Strategie\_EN\_2024.docx  
 11 ONE-PAGER\_EHDS\_-\_Kritischer\_Hintergrund\_intern\_VI-B.docx  
 12 ONE-PAGER - Digital Health Authority\_Format\_korrekt.docx  
 13 One\_Pager\_EHDS\_HBM\_2024\_06\_28.docx



# Beispiel: Software Development

- **Coding-LLMs via vLLM-API in VS Code integriert**
- LLM liest den gesamten Ordner
- Erstellt Code auf Knopfdruck, kann Änderungen auf Knopfdruck übernehmen
- Hohe Zufriedenheit unserer Entwickler
- Verschiedene Stufen wie agentisch System agiert



# Quellenangaben aus internen Dokumenten

KI X/B GPT-OSS:120b ▾ +  
Als Standard festlegen

oi KI X/B GPT-OSS:120b  
allgemein anwendbare KI mit Zugriff auf alles bisher hinzugefügte Wissen für die Gruppe  
X/B

Wie kann ich Ihnen heute helfen?

+

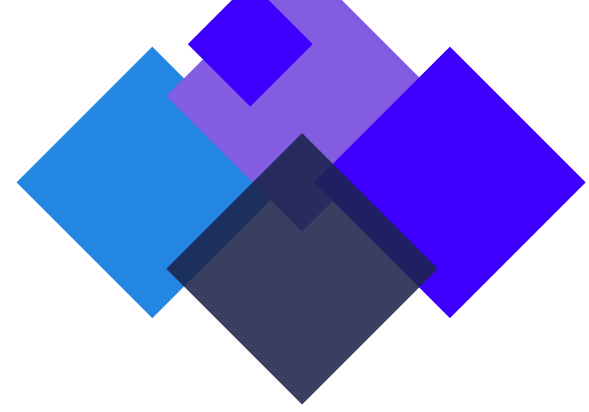
Websuche

Code-Interpreter

Give me ideas  
for what to do with my kids' art

Help me study

- **Transparenter Score der Relevanz einer Quelle**
- Bei Klick gelangt man direkt zum Original-File
- **Websuche standardmäßig aus → 100 % on premise**



05

TEIL 5

# Erfahrungen am Gesundheitsministerium

BMASGPK — Setup, Hardware, Learnings



# Vom Piloten zum Produktivsystem

- **06/2025: Beschaffung erster GPU für 700€ für ersten PoC**
- Ersteinrichtung mit Ollama
- **Überzeugende Ergebnisse des PoC**
- **Herbst 2025: Beschaffung GPU Server**
- Frühjahr 2026: Integration in Proxmox Cluster, Integration mit Stammportal (Anhebung des Stammportal auf OIDC), Einrichtung der PROD Instanzen inkl. Dokumentation
- Schulungen für User sind essenziell! Gleichzeitig Feedback der Fachabteilungen für Weiterentwicklungen



# Abdeckung & Erfahrungen

## Gut abgedeckt

- Briefing- & Textentwürfe
- Einzelne Fragen beantworten
- Wissensmanagement
- Teilw. Transkription & Übersetzung
- Software Entwicklung

## Learnings

- Hoher Mehrwert bei erstaunlich wenig Aufwand
- vLLM skalierbar, hohe Performance von MoE-LLMs
- Kleine LLMs stoßen bei Tool-Calls an Grenzen
- Spezial-Use-Cases brauchen eigene Pipelines
- Lange Texte bleiben schwierig
- Konfiguration von vLLM, RAG-Prozessen etc. nicht trivial
- Proxmox bietet eine sehr gute Basis
- Fine-tunes sind nur für spezifische Fachanwendungen nötig



# Massive Fortschritte bei LLMs

- **Gemma 4: LLM in Enterprise-Qualität auf 32 GB VRAM**
- **Speech-to-Text: Whisper Turbo, Voxtral u. a. laufen auf noch weniger VRAM**
- **Die Hardware-Anforderung wird kontinuierlich kleiner (wenige Tausend Euro)**
- **Dank neuer LLMs, Embedding Modelle etc. wird die Qualität immer besser**

**Alternative ohne Hardware-  
Investition: RAG auf eigener  
Infrastruktur, LLM in der Cloud**

**Hybrid RAG**



# Weitere praktische Open Source Lösungen



**Chat-Assistenten**

Open WebUI · RAGFlow



**Transkription**

Scriberr · aTrain · Vibe · Meetily



**Übersetzungen**

LibreTranslate



**Prozessautomatisierung**

n8n



*LibreTranslate*



# Aktuelle Grenzen

- Powerpoint Präsentationen sind technisch zu komplex (gäbe Open Source Lösung “Presenton”)
- Bildgenerierung verbraucht viel GPU-Leistung
- Qualität von on premise Coding LLMs natürlich hinter Claude Opus etc.
- Keine hochgradig agentischen Workflows
  
- **Abgedeckt werden ca. 70% der Use Cases im Arbeitsalltag**



# Ausblick & Chancen

- **Enge Zusammenarbeit mit Fachabteilungen, optimierte RAG-Pipelines**
- **Rule as Code: Prüfregele aus Regulatorik extrahieren, fachlich gesichert, KI-gestützte Antragsprüfung**
- **Entwicklung von mehr MCP Toolservern**
- **On-Prem Meeting-Transkription & Integration mit Knowledge Bases**
- Integration in Fachanwendungen
- Prozessautomatisierungen und Chat-Funktionen auf Websites
- Hochrisiko-Bereiche gemäß AI Act vermeiden



# Rule as Code zur Prozessunterstützung





# Vorteile von on premise Open Source KI-Lösungen



Keine Abhängigkeit von IT-Dienstleistern



Viel günstiger als Nutzer-Lizenzen (auch APIs)



Einfacher Wechsel auf andere Provider



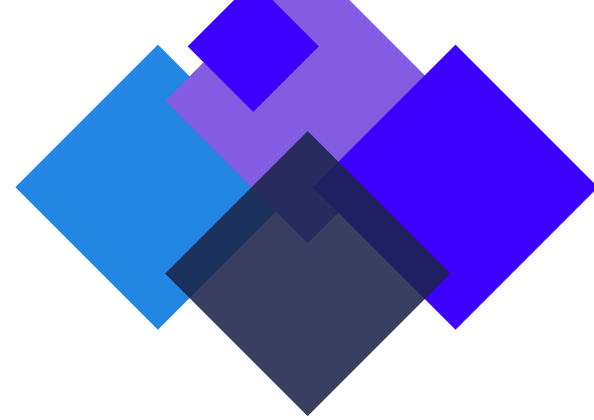
Hardware amortisiert sich schnell



Keine Datenschutz-Bedenken



Integration in bestehende Prozesse



# Danke für die Aufmerksamkeit!

## ◆ Bleiben wir in Kontakt

Philipp Schardax, MSc. M.A.

[philipp.schardax@onpremises.org](mailto:philipp.schardax@onpremises.org)

[onpremises.org](https://onpremises.org)

LinkedIn: [onpremises.org](https://onpremises.org)

LinkedIn:

