Open-Source und Offline-KI

ki.
erlernen.
jetzt

Beruflicher Hintergrund

Profil
Johannes Hofmann, selbstständiger Art Director und KI-Manager (IHK), mit jahrelanger nationaler und internationaler Agenturerfahrung und 17 Jahren Selbstständigkeit

Digital-Expertise und KI
- Einführung des ersten digitalen Schauraums der Audi AG, Prototypen für Content-Module, Technische Spezifikationen
- Prototypen für immersive Gamification-Module im Audi Forum Ingolstadt
- Content für das lokale Digital-Signage-System des Audi Forum Ingolstadt und Einbindung von Echtzeitdaten und Sprachadaptionen von Video-Content
- UX/UI-Design für Augmented Reality-Formate für Audi-Technik
- KI-gestützte Katalogautomatisierung
- Produktivsystem eines Corporate-Voice-LLM für eine deutsche Automotive-Agentur

ki.
erlernen.
jetzt

"Wer seine Werkzeuge nicht kontrolliert, wird von ihnen kontrolliert."

— Open-Source-Prinzip

ki.
erlernen.
jetzt

Verstehen – Ausprobieren – Reflektieren

1. Das Problem: Warum digitale Souveränität?
2. Die Lösung: Open-Source KI lokal nutzen
3. Live-Demo: OpenWebUI in Aktion
4. Praxis: Was brauche ich wirklich?
5. Next Steps: Heute Abend loslegen

ki.
erlernen.
jetzt

Kann ein Staat souverän bleiben, wenn seine KI-Modelle aus dem Ausland kommen?

ki.
erlernen.
jetzt

Was passiert, wenn ChatGPT morgen offline geht?

Rechtliche Gründe? Datenschutz-Urteil?
Server-Ausfall? Ihre Pressetexte, Förderanträge,
Content-Planung...
...alles steht still.

Digitale Souveränität = Handlungsfähig bleiben

ki.
erlernen.
jetzt

Drei unbequeme Wahrheiten über Cloud-KI

1. Ihre Prompts trainieren fremde Modelle
   → Steht in den AGBs

2. Token-Limits & Kosten
   → 1 Mio. Tokens = 15€

3. Keine Internetverbindung = keine KI

ki.
erlernen.
jetzt

Cloud-KI vs. Lokale KI: Der ehrliche Vergleich

Cloud-KI: Schnell, bequem, abhängig

- Pro: Beste Modelle, keine Installation
- Contra: Kosten, Datenschutz

Lokale KI: Langsamer, souverän

- Pro: Geringe Kosten, volle Kontrolle
- Contra: Hardware, Setup

ki.
erlernen.
jetzt

Die Lösung: Open-Source-Modelle lokal nutzen

Llama, Mistral, Phi
→ Leistungsstarke Modelle, frei verfügbar

OpenWebUI, LM Studio
→ Tools mit Nutzerfreundlichkeit

Einmal installiert: Unbegrenzt nutzbar

ki.
erlernen.
jetzt

Der Automation-Vorteil: Latenz wird irrelevant

Früher: 4 Tage für 10 Pressetexte

Heute: 2 Stunden Arbeitszeit

→ n8n-Workflow läuft nachts

→ Morgens: Beste auswählen

Die Maschine arbeitet 24/7. Sie kuratieren.

ki.
erlernen.
jetzt

Live-Demo: Lokale KI in Aktion

Was Sie gleich sehen:

→ OpenWebUI läuft auf diesem Laptop
→ Kein Internet, keine Cloud
→ Prompt: Pressetext
→ Side-by-Side: Cloud vs. Lokal

ki.
erlernen.
jetzt

Tool 1: OpenWebUI – Die All-in-One-Lösung

Was ist OpenWebUI?

- ChatGPT-ähnliche Oberfläche
- Unterstützt Llama, Mistral, Phi
- Browser: localhost:8080

Für wen? Fortgeschrittene, Teams

ki.
erlernen.
jetzt

Tool 2: LM Studio – Der Einsteiger-Favorit

Was ist LM Studio?

- Desktop-App (Win/Mac/Linux)
- Ein-Klick-Installation
- Perfekt für erste Schritte

Für wen? Einsteiger

ki.
erlernen.
jetzt

Was brauche ich? Hardware-Anforderungen

Use Case 1: Einfache Texte
→ Llama 3.2 3B | 8 GB RAM

Use Case 2: Pressetexte
→ Llama 3.1 8B | 16 GB RAM

Use Case 3: Komplexe Analysen
→ Mixtral 8x7B | 32 GB RAM

ki.
erlernen.
jetzt

1.Lokale LLM-Nutzung für Text- und Bild-generierung auf Laptop/Desktop-PC/Mac

- Empfohlene Hardware: CPU: Moderne Mehrkern-Prozessoren wie Intel Core i7 oder AMD Ryzen 7
- RAM: Mindestens 32 GB
- GPU: Dedizierte Grafikkarte mit mindestens 12 GB VRAM, z.B. NVIDIA RTX 3060
- Geschätzte Kosten: Desktop-PC: Zwischen 1.500€ und 2.500€, abhängig von den spezifischen Komponenten Laptop: Hochleistungs-Laptops mit vergleichbarer Ausstattung liegen zwischen 2.000€ und 3.000€ · Mac mini: 3.500 € · Mac Book Pro: 2000 – 4000 €

ki.
erlernen.
jetzt

2. Lokale LLM-Nutzung für kleine Arbeitsgruppen (5–15 Personen) inklusive Automatisierungen

- Empfohlene Hardware: Leistungsstarke Mehrkern-Prozessoren wie AMD Ryzen 9
- RAM: Mindestens 128 GB
- GPU: NVIDIA RTX 4090 mit 24 GB VRAM
- Netzwerk: Gigabit-Ethernet
- Geschätzte Kosten: Desktop-PC: Zwischen 3000 € und 5000 €, abhängig von den spezifischen Komponenten

ki.
erlernen.
jetzt

3. Lokale LLM-Nutzung in einem KMU mit ca. 300 Mitarbeitern

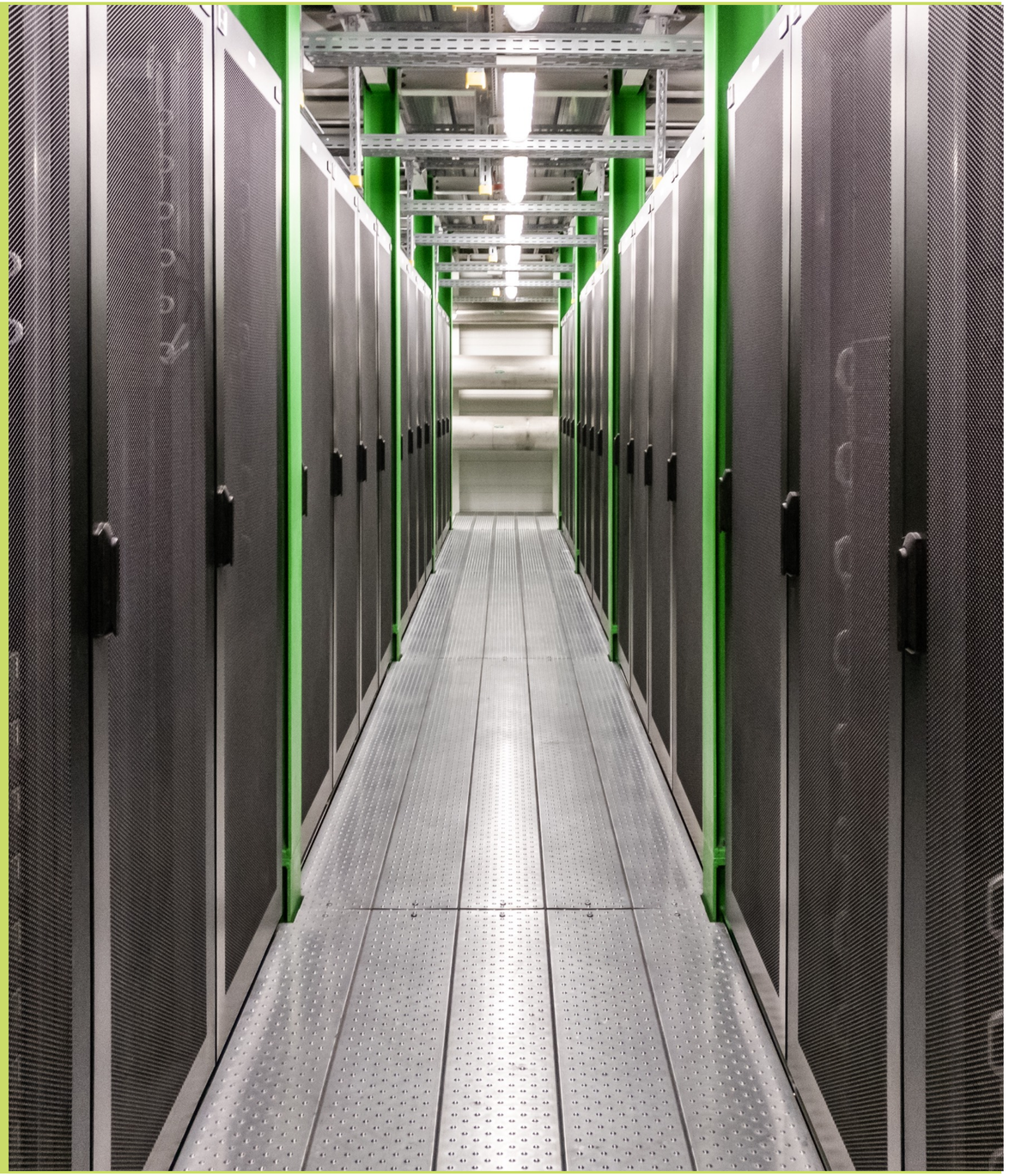
- Empfohlene Hardware: Mehrere Server mit Intel Xeon Gold oder AMD EPYC Prozessoren
- RAM: Pro Server mindestens 256 GB
- GPU: Mehrere NVIDIA A100 oder H100 GPUs mit jeweils mindestens 80 GB VRAM
- Speicher: Schnelle NVMe-SSDs und redundante HDDs/SSDs
- Netzwerk: Gigabit-Ethernet
- Geschätzte Kosten: Pro Server: Zwischen 15.000€ und 30.000€, abhängig von der Konfiguration. Gesamtkosten: Für ein Cluster von 3–5 Servern liegen die Investitionen zwischen 45.000€ und 150.000€

ki.
erlernen.
jetzt

4. Enterprise-Lösung für 1.000 Mitarbeiter und mehr

- Empfohlene Hardware: High-End-Server mit Multi-Socket-Motherboards, ausgestattet mit Intel Xeon Platinum oder AMD EPYC Prozessoren
- RAM: Pro Server mindestens 1 TB
- GPU: Mehrere NVIDIA A100 oder H100 GPUs mit jeweils mindestens 80 GB VRAM
- Speicher: Enterprise-Grade NVMe-SSDs und umfangreiche, redundante Speicherlösungen
- Netzwerk: 40-Gigabit-Ethernet oder höher
- Geschätzte Kosten: Pro Server: Zwischen 50.000€ und 100.000€, abhängig von der Konfiguration. Gesamtkosten: Für ein Cluster von 3–5 Servern liegen die Investitionen zwischen 500.000€ und 2.000.000€

ki.
erlernen.
jetzt



Praxis-Beispiel: Content-Erstellung

Aufgabe: Pressetext für Ausstellung

Prompt-Struktur:

- Wer, Was, Wann, Wo
- Zielgruppe
- Tonalität

Ergebnis: Entwurf in 30-60 Sek

ki.
erlernen.
jetzt

Was funktioniert NICHT gut lokal?

❌ Sehr aktuelle Informationen

❌ Multimodale Aufgaben (Bild+Text)

❌ Team-Kollaboration

💡 Hybrid-Ansatz für Spezialfälle

ki.
erlernen.
jetzt

Installation in 3 Schritten

Schritt 1: Download
→ lmstudio.ai (~200 MB)

Schritt 2: Modell laden
→ "Llama 3.2 3B" (~2 GB)

Schritt 3: Loslegen
→ Chat → Prompt → Enter

ki.
erlernen.
jetzt

Ihre nächsten Schritte

- Heute Abend: LM Studio installieren
- Diese Woche: Erstes Modell testen
- Nächsten Monat: Workflow automatisieren

 Handout liegt aus!

ki.
erlernen.
jetzt

Lust auf selbstbestimmtes Handeln?

Heute Abend: Installation starten

Kontakt:

Johannes Hofmann
info@ki-erlernen.jetzt
Hamburg
https://ki-erlernen.jetzt

Link zum Handout

💡 Handout mitnehmen!

ki.
erlernen.
jetzt

Industrialisierung vs. KI-Revolution: Ähnlichkeiten und Unterschiede

- **Die Industrialisierung** revolutionierte die Produktion von Waren, während die generative KI **Revolution in der kreativen Industrie einsetzt**.

- **Automatisierung kreativer Prozesse**: Industrialisierung ermöglichte die Massenproduktion durch Maschinen. Generative KI-Tools und -Technologien ermöglichen die Automatisierung von kreativen Prozessen, wie z.B. der Erstellung von Inhalten.

- **Demokratisierung der Produktion:** Die Industrialisierung machte Produkte für eine breitere Bevölkerungsschicht zugänglich. Generative KI-Tools und -Technologien machen es auch Nichtexperten ermöglichen, kreative Inhalte zu erstellen, wird quasi demokratisiert.

ki.
erlernen.
jetzt

Lokale Software zur Nutzung der Modelle (nicht vollständig)

Programm	Mac OSX	PC Windows
LM Studio	Ja	Ja
GPT4All	Ja	Ja
Open WebUI (über Terminal, Docker) Besonders für Arbeitsgruppen geeignet Browser-basiert	Ja	Ja

ki.
erlernen.
jetzt

Lokale Software zur Nutzung der Modelle (nicht vollständig)

Programm	Mac OSX	PC Windows
Diffusion Bee, leicht installierbar über Installer	Ja	Nein
NMKD, leicht installierbar über Installer	Nein	Ja
Pinokio, leicht installierbar über Installer	Ja	Ja
Automatic1111, im Terminal installierbar	Ja	Ja
Fooocus, im Terminal installierbar	Ja	Ja
ComfyUI, im Terminal installierbar	Ja	Ja

ki.
erlernen.
jetzt