AI ENGINEERING

KOSTENLOSE CHECKLISTE

10 Schritte zum eigenen AI-Stack

DSGVO-konform · 100% lokal · EU AI Act ready

🔒 Keine Cloud — deine Daten bleiben lokal

💶 Break-even in ~5 Monaten

⚡ In 2–4 Stunden einsatzbereit

Warum ein eigener AI-Stack?

Cloud-AI (ChatGPT, Copilot, Claude API) ist praktisch — aber **deine Daten verlassen dein Haus**. Kundendaten, interne Dokumente, Strategiepapiere landen auf US-Servern.

Mit einem eigenen AI-Stack behältst du die Kontrolle. **Kein SaaS-Abo. Keine DSGVO-Risiken. Kein Lock-in.** Diese Checkliste zeigt dir den direkten Weg — ohne Vorkenntnisse.

01

Hardware planen

- [ ] Server mit min. 16 GB RAM, 4 CPU-Kerne
- [ ] GPU empfohlen: NVIDIA RTX 3060+ (6 GB VRAM)
- [ ] Alternative: gebrauchter PC (~300–600 €)
- [ ] Gigabit-LAN (kein WLAN für AI-Server)

💡 Alter Büro-PC + SSD reicht für DSGVO-Compliance. Keine Cloud nötig.

02

Betriebssystem & Basis-Setup

- [ ] Ubuntu 22.04 LTS oder Debian 12
- [ ] SSH Key-basiert (kein Passwort-Login!)
- [ ] Firewall aktivieren: ufw enable
- [ ] Docker + Docker Compose installiert

03

Lokales LLM mit Ollama

- [ ] Ollama installieren (1 curl-Befehl)
- [ ] Modell laden: ollama pull llama3.2:3b (2 GB)
- [ ] Test: Frage stellen, Antwort erhalten
- [ ] API auf Port 11434 verfügbar

💡 llama3.2:3b für Start, llama3.1:8b mit GPU für bessere Qualität.

04

Chat-UI: Open WebUI

- [ ] Open WebUI via Docker starten
- [ ] Browser öffnen: localhost:3000
- [ ] Admin-Account anlegen
- [ ] Alle Ollama-Modelle sofort verfügbar

💡 Sieht aus wie ChatGPT — deine Nutzer brauchen keine Einschulung.

05

Workflow-Automation mit n8n

- [ ] n8n via Docker starten (Port 5678)
- [ ] Ollama-Verbindung konfigurieren
- [ ] Erster Workflow: E-Mails mit KI zusammenfassen
- [ ] Webhooks für externe Trigger

💡 Use Case: Berichte generieren, E-Mails sortieren, Dokumente analysieren.

06

RAG: Deine Dokumente mit KI

- [ ] ChromaDB oder Qdrant als Container
- [ ] PDFs, Word-Dateien, Wikis indexieren
- [ ] RAG-Pipeline in Open WebUI aktivieren
- [ ] Test: Fragen zu eigenen Dokumenten

💡 RAG = LLM beantwortet Fragen aus deinen eigenen Dokumenten — ohne Cloud.

07

Monitoring & Alerting

- [ ] Prometheus + Grafana deployen
- [ ] Node Exporter auf allen Servern
- [ ] Alerts: CPU >85%, RAM >90%, Disk >80%
- [ ] Alert via E-Mail, Slack oder Mattermost

08

DSGVO dokumentieren

- [ ] Verarbeitungsverzeichnis (Art. 30) anlegen
- [ ] Technische Maßnahmen (TOMs) notieren
- [ ] AVV mit externen Diensten prüfen
- [ ] Datenschutzerklärung auf Website aktuell?

💡 Kein Datentransfer in USA → kein SCCs-Problem. Eigener Stack = DSGVO-Vorsprung.

09

Security & Zugriffskontrolle

- [ ] Admin-Oberflächen hinter VPN (WireGuard)
- [ ] Starke Passwörter + 2FA überall
- [ ] API-Keys quartalsweise rotieren
- [ ] Backup: Wiederherstellung 1× testen!

10

Skalierung & Optimierung

- [ ] Multi-Node Setup: Docker Swarm
- [ ] GPU-Sharing zwischen Services
- [ ] Modell-Quantisierung: Q4_K_M = 40% weniger RAM
- [ ] Kosten-Review nach 6 Monaten

💶 Kosten-Vergleich: Eigener Stack vs. Cloud-APIs

Setup	Hardware (einmalig)	Strom/Jahr	Cloud-Äquivalent/Jahr	Break-even
Minimal (CPU only)	~300 €	~50 €	~600 €	**~6 Monate**
Standard (RTX 3060)	~800 €	~120 €	~2.400 €	**~5 Monate**
Produktion (RTX 4090)	~2.500 €	~300 €	~12.000 €	**~3 Monate**

AI ENGINEERING

PLAYBOOK — 8 KAPITEL

Vom leeren Server zum produktiven AI-Stack in 7 Tagen

Diese Checkliste ist ein Vorgeschmack. Das vollständige Playbook führt dich Schritt für Schritt durch den gesamten Aufbau.

PLAYBOOK · 8 KAPITEL

Der Lokale AI-Stack

- ✓ 70+ Seiten Schritt-für-Schritt Anleitung
- ✓ Docker Swarm Multi-Node Setup mit echten Konfigurations-Files
- ✓ n8n: 13 fertige, produktionsreife Workflows
- ✓ Grafana: 22 Monitoring-Dashboards (Import-ready)
- ✓ DSGVO Art. 30 Template — sofort ausfüllbar
- ✓ Alle Scripts & Befehle — copy-paste ready

49 EUR · einmalig · sofortiger Download

Jetzt kaufen → ai-engineering.at

Fragen? kontakt@ai-engineering.at