

LLM-Hosting in Pflegediensten

Wie können generative KI-Systeme datenschutzkonform auf eigenen Servern als Chatbots für Beratungsanfragen betrieben werden?

KI in der Pflegeberatung

Dr. Lothar Hotz

Hamburger Informatik Technologie-Center e.V.

28. Januar 2026

lothar.hotz@hitec-hamburg.de



Hamburger Informatik
Technologie – Center e.V.



University of Hamburg

Übersicht

- LLMs, Cloud-Act, KI-Landschaft
- Entscheidungskriterien
- Hosting-Möglichkeiten für Pflegedienste

Was sind lokale LLMs?

- Definition
 - Große Sprachmodelle, Large Language Models (LLMs): KI-Systeme, die auf riesigen Textmengen trainiert sind, um Sprache zu verstehen und zu erzeugen
 - Lokale LLMs: laufen vollständig auf eigenen Servern oder Geräten – keine externe Cloud erforderlich

Was sind lokale LLMs? (2)

- Wie LLMs in Pflegeeinrichtungen helfen können
 - Im Büro: Berichte, E-Mails und Dokumentation verfassen
 - Pflegeleitlinien oder Akten durchsuchen, zusammenfassen und analysieren
 - Sichere, lokale Unterstützung für tägliche Arbeitsabläufe ermöglichen
 - Usw.

Was sind lokale LLMs? (3)

- Wesentliche Merkmale
 - Modelle wie Llama 3, Mistral, Gemma, Phi-3 (lokal betreibbar)
 - Open Source und anpassbar
 - Unterschiedliche Größen → passend zur verfügbaren Hardware (7B* – 30B Parameter üblich, bis 600B möglich)
 - Feinabstimmung mit internen Dokumenten oder Pflegeleitlinien möglich
- Unterschied zu Cloud-LLMs
 - Cloud-LLMs (z. B. ChatGPT, Gemini, Claude) → von US-Anbietern gehostet
 - Lokale LLMs → vollständig unter Kontrolle der Einrichtung, offline-fähig, datenschutzsicher

Warum lokale LLMs

- Vertrauen & Datenschutz
 - Patientendaten = hochsensibel
 - DSGVO-Einhaltung essenziell
 - Volle lokale Datenkontrolle – kein externer Zugriff
- Cloud Act*
 - US-Regierung kann jedes US-Unternehmen auffordern, einen vollen Zugriff auf Daten zu gewähren.
 - Egal, wo die Daten liegen.
 - Davor schützen: Alle Daten vorher verschlüsseln, dann aber nur *Cloud-Speicher*

*https://de.wikipedia.org/wiki/CLOUD_Act

KI-Landschaft

- US-Anbieter
 - Microsoft / OpenAI, Google, Anthropic, Meta, Amazon
 - Cloud Act (USA) → potenzieller Regierungszugriff auf Daten
 - Hohe Qualität, geringe Souveränität
- Chinesische Anbieter
 - DeepSeek, Alibaba Qwen, Baidu Ernie
 - Lokaler Einsatz möglich, aber politisch beeinflusst, geringe Transparenz

KI-Landschaft (2)

- Europäische / Lokale Modelle
 - Mistral (FR), Aleph Alpha (DE), Apertus (CH), Open Assistant / Hugging Face (EU)
 - DSGVO- / EU-AI-Act-konform
 - On-Premise oder private Cloud-Hosting → Datensouveränität & Vertrauen
- Fazit:
 - USA = Datenschutzrisiko
 - China = Intransparent
 - EU = Sicher & ethisch

Übersicht

- LLMs, Cloud-Act, KI-Landschaft
- **Entscheidungskriterien**
- Hosting-Möglichkeiten für Pflegedienste

Hosting & Zuverlässigkeit

- Vorteile von EU- & lokalem Hosting
 - Betrieb nach EU-Recht (DSGVO, EU AI Act)
 - Transparent und ethisch ausgerichtet
 - Besseres Vertrauen bei Patienten & Mitarbeitenden
- Abhängigkeit & Zuverlässigkeit
 - Keine Abhängigkeit von GAFAM* / US-Clouds
 - Offline-Fähigkeit für Pflegeeinrichtungen
 - Stabiler Betrieb unabhängig von Politik oder Abos

* Tech-Giganten **G**oogle (Alphabet), **A**mazons, **F**acebook (jetzt Meta), **A**pple und **M**icrosoft

Kosten & Kontrolle

- Kosten & Nachhaltigkeit
 - Keine API- oder Tokengebühren
 - Nutzung vorhandener Hardware
 - Planbare Betriebskosten
- Modellkontrolle & Relevanz
 - Lokales Fine-Tuning mit Pflege-Terminologie
 - Jederzeit aktualisierbar (keine veralteten ChatGPT-Versionen)
 - Open Source Modelle nutzbar, auch die aus China!
 - Maßgeschneiderte Apps → ChatBOT passt zu realen Aufgaben
 - Datenschutz vs. sehr große Modelle der GAFAM

Qualität & Validierung

- Abhängig von Trainingsdaten & Feinabstimmungsqualität
- Validierungs- oder Prüfschritte für Pflegeempfehlungen einbauen
- Antworten an verifizierten Quellen / Leitlinien prüfen
- Regelmäßige Tests auf Bias & veraltete Inhalte

Übersicht

- LLMs, Cloud-Act, KI-Landschaft
- Entscheidungskriterien
- **Hosting-Möglichkeiten für Pflegedienste**

Lokale LLMs in Pflegeeinrichtungen einführen

- Notwendigkeit individueller Entwicklung
 - Kein einheitlicher Chatbot passt für alle Pflegesituationen
 - Lokale Modelle müssen mit Einrichtungsdaten (z. B. Dokumentation, Prozess) integriert werden
 - Individuelle Apps machen den Chatbot praxisgerecht und datensicher
 - Lösungen wie fobizz* zeigen, wie datensichere Frontends für KI im Bildungsbereich funktionieren
 - Ähnliches Konzept für Pflege denkbar, aber Datenabfluss nach US möglich
 - Ziel: Einfacher, datenschutzkonformer ChatBOT, zugeschnitten auf den Bedarf des Pflegepersonals

EU-LLM-Hoster

- EU-LLM-Anbieter in einer EU-Cloud
- chat.mistral.ai
- MS/Azure, aber Cloud-Act
- Token-Kosten

On-Premise

- Modell läuft auf internem Server oder vorhandener Einrichtungshardware
- Volle Datenkontrolle, vollständig DSGVO-konform
- Funktioniert offline, unabhängig von externen Anbietern
- Erfordert lokale IT-Wartung, Updates und Backups
- Hardwareanforderung: moderne GPU oder leistungsstarke CPU

Private EU-Cloud

- Hosting in zertifiziertem deutschem/EU-Rechenzentrum
- Daten bleiben innerhalb der EU-Rechtsprechung
- Skalierbare Leistung ohne lokale Hardware-Investition
- Betrieb unter strengen Auftragsverarbeitungsvereinbarungen
- Stabile Internetverbindung & vertrauenswürdiger Provider notwendig

Kosten lokaler LLMs

- On-Prem-GPUs sind teuer (€2.000 – € 6.000 + Strom)
 - EU-LLM-Hoster = keine Hardware, aber Tokengebühren summieren sich
 - Abwägung: Hardwareinvestition vs. nutzungsabhängige Kosten
 - Eigene Server mehrere €10.000
-
- Entscheidung: On-Prem vs. Server auf EU-Cloud vs. EU-LLM-Hoster

Notwendige Zusammenarbeit

- Einzelne Pflegeeinrichtungen haben begrenzte Ressourcen für Hosting und Entwicklung
- Kooperation mit IT-Anbietern oder öffentlichen Rechenzentren
- Gemeinsamer Betrieb sichert Compliance und Bezahlbarkeit
- Vorschlag: Zentrale Ressourcen für mehrere Pflegedienste bereitstellen

Vielen Dank für das Interesse!

Get in touch

Hamburger Informatik Technologie-Center e.V.

HITEC e.V.

Mail: lothar.hotz@hitec-hamburg.de

Vogt-Kölln-Straße 30

ab 23.3.: Bundesstraße 56 b

D-22527 Hamburg

