Lokale KI-Modelle: Das unterschätzte Potenzial abseits der Cloud

Paul Mielnikowski (using ChatGPT free)

February 2025

1 Einleitung

In der aktuellen Debatte um Datenschutz und KI-Regulierung steht fast ausschließlich die Cloud-Nutzung im Fokus. Kaum jemand spricht darüber, dass leistungsstarke KI-Modelle wie DeepSeek oder LLaMA lokal betrieben werden können – ohne Daten in fremde Server zu senden. Diese technische Möglichkeit bleibt politisch weitgehend unbeachtet.

2 Lokale KI: Was ist möglich?

Dank Plattformen wie Ollama oder LM Studio können Nutzer leistungsfähige Sprachmodelle mit Hardware im Wert von wenigen tausend Euro betreiben. Modelle wie DeepSeek, Mixtral oder LLaMA 3 lassen sich problemlos lokal installieren und nutzen – mit voller Datenkontrolle.

3 Politische Wahrnehmung

Weltweit konzentrieren sich Regierungen auf Cloud-KI-Dienste. Selbst in Deutschland, wo Datenschutz zentral ist, richtet sich die Diskussion primär gegen Plattformen wie ChatGPT oder DeepSeek Cloud. Die lokale Nutzung offener Modelle bleibt bisher außerhalb des politischen Radars.

4 Rechtliche Situation

Laut der aktuellen EU-KI-Verordnung (AI Act) fallen Open-Source-Modelle grundsätzlich nicht unter strenge Auflagen, es sei denn, sie werden für Hochrisikoanwendungen genutzt.
Die deutsche Datenschutzgesetzgebung richtet sich primär gegen Cloud-basierte Datenübertragungen.

5 Implikationen

5.1 Datenschutz

Lokale Nutzung schützt sensible Daten und vermeidet Cloud-Risiken.

5.2 Souveränität

Firmen und Institutionen können unabhängig agieren.

5.3 Sicherheitsrisiken

Fehlender politischer Fokus könnte zur unkontrollierten Nutzung führen.

6 Kostenrechnung für den lokalen Betrieb von KI-Modellen

6.1 Hardware-Kosten

Um ein KI-Modell wie LLaMA, DeepSeek oder Mixtral lokal auszuführen, braucht man eine solide Recheninfrastruktur. Die Grundausstattung für einen leistungsstarken PC oder Server könnte wie folgt aussehen:

- **NVIDIA GPU (z. B. RTX 3090 oder A100)**: 1.500 bis 3.000 Euro (je nach Modell)
- **Prozessor (AMD Ryzen 9 oder Intel Xeon)**: 500 bis 1.000 Euro
- **Arbeitsspeicher (64 GB RAM)**: ca. 300 bis 500 Euro
- **Speicher (SSD, 1-2 TB)**: ca. 150 bis 300 Euro
- **Netzteil, Kühlung, Gehäuse**: ca. 200 bis 400 Euro

Insgesamt würde eine einmalige Anschaffung für eine Maschine, die leistungsstark genug für KI-Modelle ist, zwischen 2.500 und 5.000 Euro liegen. Diese Hardware ist dann in der Lage, viele der modernen Open-Source-Modelle effizient auszuführen.

6.2 Stromkosten

Der Stromverbrauch hängt stark von der Rechenleistung und der Nutzungsdauer ab. Eine leistungsstarke GPU wie die NVIDIA RTX 3090 verbraucht im Dauerbetrieb etwa 300 Watt. Wenn das System rund um die Uhr läuft (24/7), sieht die Rechnung wie folgt aus:

- Stromverbrauch pro Stunde: 0,3 kWh
- Stromverbrauch pro Tag (24 Stunden): 7,2 kWh
- Stromverbrauch pro Monat (30 Tage): 216 kWh

Bei einem Strompreis von 0,30 €/kWh ergibt sich folgendes:

- Stromkosten pro Tag: 2,16 Euro
- Stromkosten pro Monat: 64,80 Euro
- Stromkosten pro Jahr: 777,60 Euro

6.3 Vergleich zu Cloud-Diensten

Cloud-Dienste wie ChatGPT Pro oder Google Cloud AI bieten Modelle, die auf monatlicher Basis abgerechnet werden. Schauen wir uns das ChatGPT Business Modell an, das etwa 200 US-Dollar pro Monat kostet.

- Kosten pro Monat: 200 USD
- Kosten pro Jahr: 2400 USD

6.3 Vergleich zu Cloud-Diensten Cloud-Dienste wie ChatGPT Business oder Google Cloud AI bieten Modelle, die auf monatlicher Basis abgerechnet werden. Das ChatGPT Business-Modell kostet etwa 200 US-Dollar pro Monat. Allerdings zeigt sich, dass OpenAI mit diesem 200-Dollar-Plan wahrscheinlich Verluste macht, da die Betriebskosten für die Serverinfrastruktur, die ständige Weiterentwicklung des Modells sowie die hohe Rechenleistung für die Nutzung der KI weit über den monatlichen Einnahmen von 200 USD liegen. Infolgedessen müsste dieser Plan eigentlich deutlich teurer sein, um die Kosten zu decken und profitabel zu bleiben.

Kosten pro Monat (ChatGPT Business): 200 USD Zusätzliche API-Kosten (je nach Nutzung): variabel, abhängig von der Token-Anzahl Kosten pro Jahr (ChatGPT Business ohne API-Nutzung): 2.400 USD (plus API-Kosten) Die Kosten für den Cloud-Betrieb sind im Vergleich zum lokalen Betrieb deutlich höher. Zudem führt die hohe Rechenleistung und Infrastruktur, die für den Betrieb eines solchen Modells erforderlich ist, dazu, dass der Cloud-Service von OpenAI unter den aktuellen Preisen mit dem 200-Dollar-Plan nicht kostendeckend ist. Das bedeutet, dass der Betrieb über die Cloud bei intensiver Nutzung oder mehreren Nutzern langfristig teurer und weniger nachhaltig ist im Vergleich zum lokalen Betrieb von Open-Source-KI-Modellen.

Die Kosten für den Cloud-Betrieb sind im Vergleich zum lokalen Betrieb deutlich höher, vor allem wenn man von mehreren Nutzern oder intensiver Nutzung ausgeht.

6.4 Langfristige Kostenersparnis

Wenn wir also die einmaligen Hardwarekosten (2.500 bis 5.000 Euro) und die Stromkosten (ca. 780 Euro pro Jahr) für den lokalen Betrieb eines KI-Modells betrachten, sind diese langfristigen Kosten weit günstiger als die regelmäßigen Cloud-Abonnements. Ein Unternehmen, das zum Beispiel ein Jahr lang KI-Modelle lokal betreibt, würde die Hardwarekosten (sofern bereits vorhanden) über einen längeren Zeitraum amortisieren und nach einem Jahr vor allem die Stromkosten tragen.

7 Fazit

Die Rechnung spricht für sich: Der Betrieb von Open-Source-KI-Modellen auf lokal verfügbaren Maschinen ist deutlich kostengünstiger als die Nutzung von teuren Cloud-Diensten. Politiker und Entscheidungsträger müssen endlich erkennen, dass die Kosten der Cloud-Nutzung nicht nur schnell steigen, sondern auch die Datensicherheit gefährden können. Lokale Open-Source-Modelle bieten eine kostengünstige, datensichere und unabhängige Alternative, die in vielen Fällen sowohl für Unternehmen als auch für Privatpersonen eine weitaus bessere Option darstellt.