6 Tipps: So betreiben Unternehmen KI-Sprachmodelle auf eigener Hardware

Utl.: Der Speicher entscheidet, ob ein Modell läuft – der Arbeitsablauf, wie schnell es ist

(Wien, 09.11.2026) Mit der passenden Hardware betreiben Unternehmen, Kanzleien und Behörden KI-Sprachmodelle im eigenen Haus. Die Frage nach dem eigenen Betrieb ist aus zwei Gründen wichtig: Die Daten bleiben im eigenen Haus, und bei dauerhafter Auslastung kann die Rechnung günstiger ausfallen als bei der Nutzung über eine Programmierschnittstelle (API). Welcher Rechner der richtige ist, hängt dabei weniger von der Größe des Modells ab als vom Arbeitsablauf, und der verändert sich gerade: Eine kurze Chatfrage umfasst wenige Wörter. Ein KI-Agent liest dagegen je Arbeitsschritt 10.000 bis 30.000 Token an Suchtreffern und Dokumentauszügen; Token sind die Textbausteine, in denen Sprachmodelle rechnen. Die folgenden sechs Tipps zeigen IT-Verantwortlichen, wie sie Modell und Rechner aufeinander abstimmen, wann sich der Kauf lohnt und wo das Gerät am besten steht. Sie beruhen auf veröffentlichten Messungen, aktuellen Preisen und einem Rechenbeispiel mit 500.000 Dokumenten.

1. Zuerst das Modell wählen, dann den Rechner

Ein Sprachmodell besteht aus Milliarden gelernter Zahlenwerte, den Parametern. Damit es Texte einlesen und Antworten schreiben kann, müssen diese Werte vollständig in einem schnellen Speicher direkt beim Rechenchip liegen. Bei einer Workstation ist das der Arbeitsspeicher auf der Grafikkarte, der Grafikspeicher; beim Mac Studio und bei KI-Boxen teilen sich Prozessor und Grafikteil einen gemeinsamen Speicher. Passt das Modell nicht hinein, lässt es sich nicht sinnvoll betreiben.

Wie viel Platz nötig ist, hängt von Größe und Speicherformat ab. Ein Modell mit rund 26 Milliarden Parametern belegt in der 4-Bit-Fassung, die jeden Wert platzsparend und etwas ungenauer ablegt, etwa 16 Gigabyte Grafikspeicher, in der 8-Bit-Fassung etwa 26 Gigabyte. Zusätzlich braucht jede laufende Anfrage einen Zwischenspeicher, in dem das Modell festhält, was es bereits gelesen hat. Je länger die Texte und je mehr Anfragen gleichzeitig laufen, desto mehr Platz belegt er. Auf einer Grafikkarte mit 96 Gigabyte Grafikspeicher bleiben neben dem 8-Bit-Modell rund 70 Gigabyte für diesen Zwischenspeicher frei, auf einer Grafikkarte mit 48 Gigabyte rund 20. Welche Modelle mit offenen

searchit
by Iphos IT Solutions GmbH
Khekgasse 35
A-1230 Wien

+43 1 869 84 00
+ 43 1 869 84 00 50
info@searchit.cc
searchit.cc

Geschäftsführer:
Ing. Christoph Wendl,
Lyubomir Ivanov
UID: ATU61293157
FN 254933 b
Handelsgericht Wien

Gewichten, deren Modelldateien also öffentlich verfügbar sind, auf welcher Grafikkarte laufen und was der Bezug über eine Schnittstelle kosten würde, zeigt ein kostenloses Vergleichsportal, das unten verlinkt ist.

„*Die Frage nach dem Grafikspeicher ist die richtige erste Frage, aber eben nur die erste*“, sagt Ing. Christoph Wendl, CEO von searchit. „*Wer Hardware für den eigenen Ablauf kauft und nicht für das größte Modell, das gerade die Runde macht, bekommt mehr Leistung für sein Budget.*“

2. Die Länge der Eingaben mitrechnen

Ein Sprachmodell arbeitet in zwei Schritten, und beide stellen andere Anforderungen an die Hardware. Beim Einlesen verarbeitet es die gesamte Eingabe auf einmal. Das ist eine große Rechenaufgabe, hier zählt also die Rechenleistung des Grafikchips. Beim Schreiben erzeugt es die Antwort Token für Token und liest dafür jedes Mal seine Parameter aus dem Speicher. Hier zählt die Speicherbandbreite, also wie schnell die Daten aus dem Speicher zum Chip gelangen.

Bei einer kurzen Chatfrage dauert das Einlesen Sekundenbruchteile, gewartet wird vor allem auf das Schreiben. Bei Dokumenten und Agenten wird das Einlesen zum Engpass. Für 20.000 Token ergeben sich rechnerisch rund 0,6 Sekunden auf einer RTX PRO 6000, 1,2 Sekunden auf einer RTX PRO 5000 und grob 3 bis 12 Sekunden auf einem Mac Studio M5 Ultra. Da ein Agent je Frage mehrere solcher Schritte macht, summieren sich diese Zeiten.

3. Gleichzeitige Anfragen von Anfang an einplanen

Eine Person im Chat, zehn Kollegen mit einem Agenten und ein Stapel von 500.000 Dokumenten sind drei verschiedene Belastungen, und jede verlangt andere Hardware. Für wenige Nutzer mit kurzen Fragen genügt ein Mac Studio (6.599 Euro mit 96 Gigabyte) oder eine KI-Box der Klasse DGX Spark (4.900 bis 6.900 Euro). Beide stehen leise auf dem Schreibtisch.

Sollen viele Kolleginnen und Kollegen oder mehrere Agenten gleichzeitig arbeiten, spielt eine Workstation mit RTX PRO 6000 (rund 22.200 Euro) ihre Stärke aus. Die Software vLLM bündelt dort viele Anfragen und lässt die Grafikkarte sie gemeinsam abarbeiten: Gemessen wurden 125 gleichzeitige Anfragen mit je 18 Token pro Sekunde. Auf dem Mac hilft das Bündeln, gemessen mit der Software llama.cpp, nur beim Schreiben.

searchit
by Iphos IT Solutions GmbH
Khekgasse 35
A-1230 Wien

+43 1 869 84 00
+ 43 1 869 84 00 50
info@searchit.cc
searchit.cc

Geschäftsführer:
Ing. Christoph Wendl,
Lyubomir Ivanov
UID: ATU61293157
FN 254933 b
Handelsgericht Wien

Lange Eingaben werden dort praktisch nacheinander eingelesen, die Anfragen warten also aufeinander.

4. Benchmarks am eigenen Ablauf prüfen

Benchmarks sind standardisierte Leistungstests, und ihre Werte gelten nur für die Bedingungen des Tests. Das verbreitete Messwerkzeug llama-bench arbeitet ohne weitere Angaben mit 512 Token Eingabe, 128 Token Ausgabe und einer einzigen Anfrage. Über einen Agenten, der 20.000 Token liest und neben anderen Anfragen läuft, sagt das wenig aus. Aussagekräftig sind Werte bei der eigenen Eingabelänge und mit so vielen gleichzeitigen Anfragen wie im Betrieb.

Auch der Softwarestand zählt: Auf einem Mac Studio M3 Ultra fiel das Einlesen von 8.192 Token nach einem Update von llama.cpp von rund 2.106 auf 1.183 Token pro Sekunde, bei gleicher Hardware und gleichem Modell. Am verlässlichsten bleibt deshalb die eigene Messung mit dem geplanten Modell und typischen Eingaben.

5. Vor dem Kauf Schnittstelle und Miete durchrechnen

Ein Sprachmodell muss nicht auf eigener Hardware laufen. Über eine Programmierschnittstelle lässt es sich auch bei einem Anbieter nutzen, abgerechnet wird dann nach verarbeiteten Token. Für einen einmaligen großen Stapel ist das fast immer günstiger. Das zeigt ein Beispiel mit 500.000 Dokumenten, je 7.500 Token Eingabe und 1.500 Token Ausgabe, und dem Modell Gemma 4 26B A4B: Bei den günstigsten Angeboten mit Verarbeitung in der EU kostet der Lauf 600 bis 850 Euro, eine passende Workstation 13.500 bis 22.200 Euro.

Eigene Hardware rechnet sich, wenn sie dauerhaft arbeitet, denn sie schreibt sich auch im Leerlauf ab. Bei voller Auslastung über 36 Monate und einem Strompreis von 0,30 Euro je Kilowattstunde kosten 1.000 Dokumente auf einer RTX PRO 6000 rechnerisch rund 0,18 Euro, über die Schnittstelle 1,20 bis 1,70 Euro. Läuft die Workstation nur acht Stunden am Tag, verdreifachen sich ihre Kosten je Dokument etwa. Auch wo Daten das Haus nicht verlassen sollen, etwa bei Berufsgeheimnisträgern, ist eigene Hardware die naheliegende Wahl. Dazwischen liegt die Miete: Europäische Rechenzentren bieten eine RTX PRO 6000 ab rund 1,25 Euro je Stunde an, ohne Kauf und mit denselben Vorteilen bei gleichzeitigen Anfragen.

searchit
by Iphos IT Solutions GmbH
Khekgasse 35
A-1230 Wien

+43 1 869 84 00
+ 43 1 869 84 00 50
info@searchit.cc
searchit.cc

Geschäftsführer:
Ing. Christoph Wendl,
Lyubomir Ivanov
UID: ATU61293157
FN 254933 b
Handelsgericht Wien

6. Den Aufstellort von Beginn an mitdenken

Was eine Grafikkarte an Strom aufnimmt, gibt sie als Wärme wieder ab, und ihre Lüfter müssen diese Wärme abführen. Darum gehört die Frage nach Lärm und Abwärme vor den Kauf. Eine KI-Box mit rund 160 Watt unter Last und gemessen höchstens 37,5 dB(A) in 46 Zentimetern Abstand kann am Arbeitsplatz stehen. Eine Workstation mit RTX PRO 6000 gehört dagegen in einen Nebenraum mit Luftzufuhr: Die Karte nimmt bis rund 530 Watt auf, die Abwärme entspricht einem Heizlüfter auf kleiner Stufe, und die deutsche Arbeitsstättenregel ASR A3.7 nennt für Ein- und Zweipersonenbüros einen Planungswert von 40 dB(A).

Rund um die Uhr kostet der Strom etwa 1.970 Euro im Jahr. Die Variante Max-Q kommt mit 300 statt 600 Watt aus, kostet rund 1.180 Euro Strom und arbeitet mit Sprachmodellen gemessen 13 bis 19 Prozent langsamer. Für Läufe über mehrere Tage empfiehlt sich eine unterbrechungsfreie Stromversorgung, damit eine kurze Unterbrechung im Stromnetz nicht den gesamten Lauf abbricht.

Mehr Infos zum Beitrag „Hardware für lokale LLMs“:

https://www.searchit-enterprise-search.com/hardware-lokale-llm/

Mehr Infos zu searchit Models:

Selbst betreiben: https://models.searchit-agent.ai/selbst-betreiben/

Kostenrechner:
https://models.searchit-agent.ai/werkzeuge/kostenrechner/

Mehr Infos zum searchit AI Agent:

https://searchit-agent.ai/

Über searchit:

Mit searchit wird das Wissen einer Organisation zum Gedächtnis ihrer KI. Der searchit AI Agent kennt Ihre Datenbestände (E-Mails, Dokumente, etc.) und arbeitet mit diesen. Dadurch wird der Agent zum digitalen Mitarbeiter. Die Technologie baut auf der langjährigen Erfahrung von searchit als Enterprise-Search-Plattform auf, die von zahlreichen Kunden wie etwa Samsung, TÜV, Technisches Museum Wien und der Bayerischen Digitalbehörde eingesetzt wird. Hersteller von searchit ist die Iphos IT Solutions GmbH mit Sitz in Wien. Ihr hauseigenes AI-Lab entwickelt die KI-Funktionen der Plattform laufend weiter. Ing. Christoph Wendl leitet gemeinsam mit Lyubomir Ivanov als Chief Executive Officer das 1998

searchit
by Iphos IT Solutions GmbH
Khekgasse 35
A-1230 Wien

+43 1 869 84 00
+ 43 1 869 84 00 50
info@searchit.cc
searchit.cc

Geschäftsführer:
Ing. Christoph Wendl,
Lyubomir Ivanov
UID: ATU61293157
FN 254933 b
Handelsgericht Wien

gegründete Unternehmen.

Rückfragehinweis für Medien:

Ing. Christoph Wendl
Geschäftsführer, Iphos IT Solutions GmbH
Khekgasse 35
1230 Wien
Tel.: +43 1 869 84 00
E-Mail: info@searchit.cc

searchit
by Iphos IT Solutions GmbH
Khekgasse 35
A-1230 Wien

+43 1 869 84 00
+ 43 1 869 84 00 50
info@searchit.cc
searchit.cc

Geschäftsführer:
Ing. Christoph Wendl,
Lyubomir Ivanov
UID: ATU61293157
FN 254933 b
Handelsgericht Wien