Enterprise Strategy Group by TechTarget

WIRTSCHAFTS-WHITEPAPER

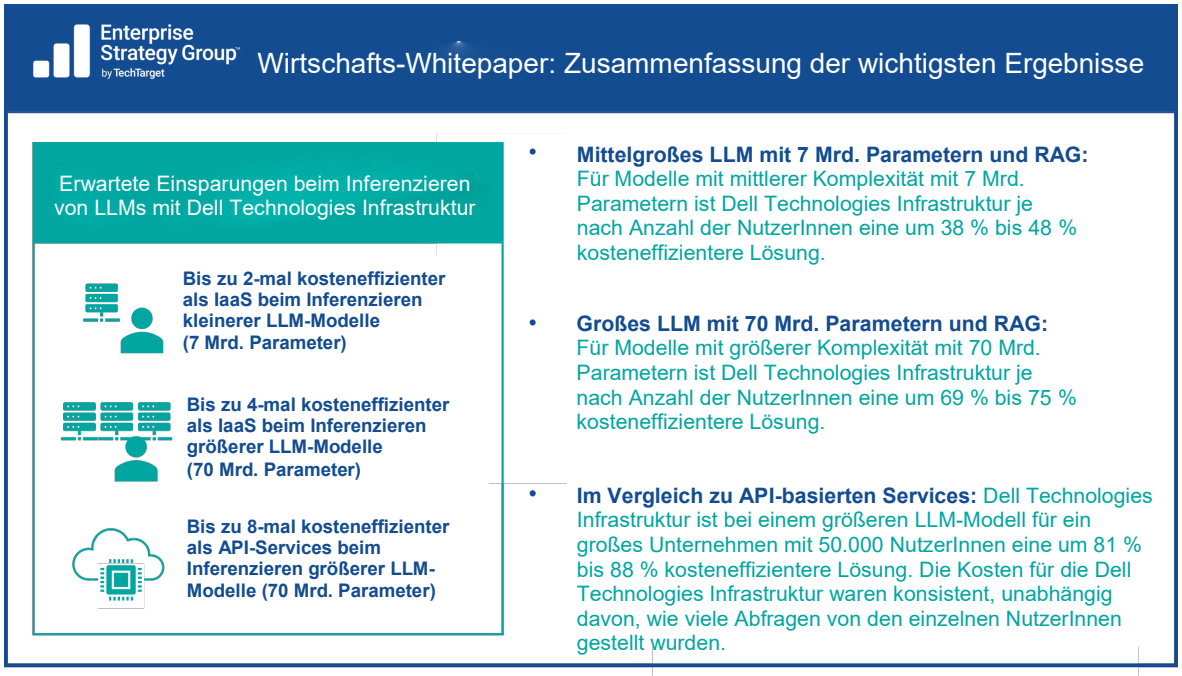
Verstehen der Gesamtkosten für das Inferenzieren großer Sprachmodelle

Wie die Nutzung von Dell Technologies On-Premise-Lösungen für das Inferenzieren von LLMs mit RAG im Vergleich zur Public Cloud oder zu tokenbasierten APIs 38 % bis 88 % kosteneffizienter sein kann

Von Aviv Kaufmann, Practice Director and Principal Validation Analyst
Enterprise Strategy Group

April 2024

Inhaltsverzeichnis



Einführung

In diesem Wirtschafts-Whitepaper werden einige der Optionen und Überlegungen zur Bereitstellung textbasierter GenAI-Funktionen (generative KI) für Unternehmen vorgestellt. Die Enterprise Strategy Group von TechTarget hat die erwarteten Kosten für das Inferenzieren großer Sprachmodelle (Large Language Models, LLMs) mit Retrieval-Augmented Generation (RAG) in einer Dell Technologies On-Premise-Infrastruktur im Vergleich zu einer nativen Infrastruktur as a Service-Lösung (IaaS) in der Public Cloud und zum OpenAI GPT-4 Turbo-LLM-Modellservice über eine API modelliert und verglichen. Wir haben festgestellt, dass Dell Technologies eine um bis zu 4-mal kosteneffizientere LLM-Inferenzierung als IaaS und eine um bis zu 8-mal kosteneffizientere Inferenzierung als die GPT-4 Turbo-API bereitstellen kann.

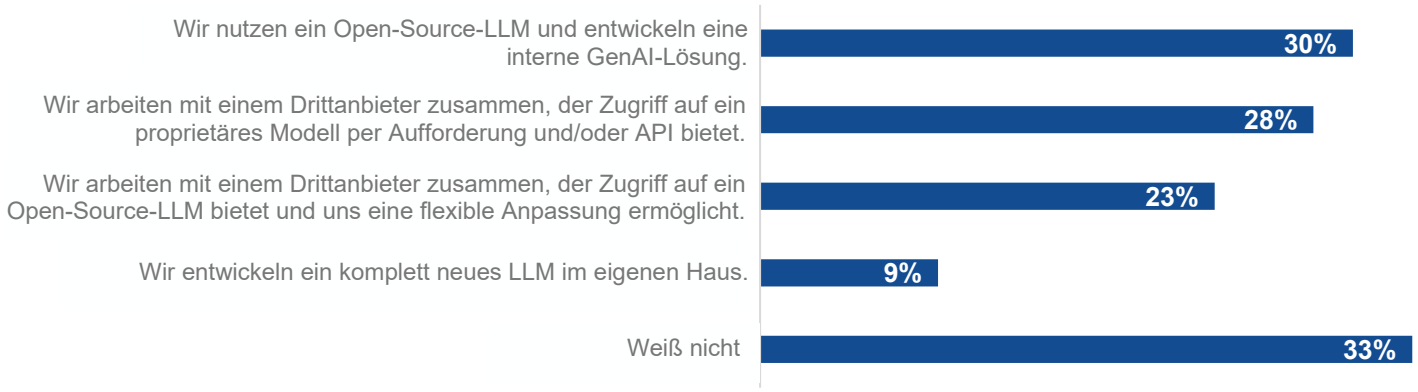
Herausforderungen

Unternehmen möchten von der Leistungsfähigkeit von GenAI und LLMs profitieren, die unternehmensspezifische Daten und anderes geistiges Eigentum nutzen, um die Erzeugung von Inhalten zu automatisieren, Fragen zu beantworten und EntscheidungsträgerInnen Erkenntnisse zur Verfügung zu stellen. Neben vielen anderen Vorteilen gaben die Befragten einer Forschungsstudie der Enterprise Strategy Group folgende primären Vorteile durch die Nutzung von GenAI in ihrem Unternehmen an: Verbesserung und/oder Automatisierung von Prozessen und Workflows, Unterstützung von Data Analytics und Business Intelligence, Steigerung der Mitarbeiterproduktivität und eine verbesserte Betriebseffizienz.[1]

Die Entwicklung von LLMs kann kostspielig und komplex sein. Unternehmen können jedoch bestehende Open-Source-LLMs mühelos erweitern, optimieren und an ihre Anforderungen anpassen. Vorgefertigte API-basierte Services wie OpenAI GPT bieten eine einfachere Lösung. Die Kosten für die Inferenzierung (d. h. das Abfragen) können sich jedoch schnell summieren, insbesondere für größere Unternehmen und komplexere LLMs. Alternativ können Unternehmen ihre eigene LLM-Inferenzierungslösung auf leistungsstarken GPU-fähigen Unternehmensservern oder gleichwertigen GPU-fähigen Cloud-Instanzen und einer Plattform für maschinelles Lernen wie AI Enterprise von NVIDIA, auf der Open-Source-LLMs ausgeführt werden, erstellen und steuern. Es überrascht nicht, dass die Enterprise Strategy Group feststellte, dass Unternehmen bei der Entwicklung und Nutzung von GenAI mit einem LLM bevorzugt auf die Strategie setzen, ein Open-Source-LLM zu nutzen und eine interne GenAI-Lösung zu entwickeln.[2]

Abbildung 1. Die meisten Unternehmen planen, ihre eigene interne GenAI-Lösung zu entwickeln

Wie wird Ihr Unternehmen generative KI entwickeln/nutzen, die von einem großen Sprachmodell (Large Language Model, LLM) unterstützt wird? (% der Befragten, N = 670, mehrere Antworten möglich)



Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

[1] Quelle: Enterprise Strategy Group-Untersuchungsbericht *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.
[2] Ebd.

Wichtige Überlegungen zur LLM-Inferenzierung

Textbasierte LLMs konzentrieren sich auf das Erlernen, Verstehen und Produzieren von textbasierten Inhalten, Antworten, Zusammenfassungen und Fragen, die auf eine bestimmte Branche, einen bestimmten Anwendungsfall und ein bestimmtes Unternehmen zugeschnitten werden können. RAG ergänzt die Ergebnisse von GenAI-Modellen mit kundenspezifischen Daten aus zusätzlichen Quellen, wodurch die Modelle genauer werden. Dies sind die am häufigsten bereitgestellten LLMs für Unternehmen. Sie können neben vielen anderen Anwendungsfällen für Chatbots, Fragen-und-Antworten-Assistenten sowie Prozessverbesserung und -automatisierung verwendet oder als Funktionen in kundenspezifische Tools und Anwendungen integriert werden. Bei der Bereitstellung von LLM-Modellen müssen Unternehmen die Infrastruktur für das Training (d. h. daten- und Compute-intensive Analysen, die für die Erstellung eines effektiven Modells erforderlich sind), die Inferenzierung (d. h. das Verarbeiten von Nutzerinteraktionen auf einem trainierten Modell) und die Feinabstimmung (d. h. das kontinuierliche Aktualisieren und Optimieren des Modells) berücksichtigen. Der Schwerpunkt dieses Berichts liegt auf der Infrastruktur, die für die Unterstützung von Inferenzierungs-Workloads erforderlich ist. Es gibt verschiedene Bereitstellungsmethoden, die für die Inferenzierung von LLMs verwendet werden können, darunter:

- **Herkömmliche Infrastruktur:** Eine gekaufte oder geleaste herkömmliche Infrastruktur, die aus Compute, Arbeitsspeicher, GPUs und Storage besteht, kann zusammen mit einer kommerziellen oder Open-Source-KI-Plattform bereitgestellt und gemanagt werden, sodass das Unternehmen die Kontrolle über alle Aspekte der Bereitstellung erhält. Bei größeren und vorhersehbaren Workloads ist diese Methode möglicherweise die kosteneffizienteste.
- **Public-Cloud-IaaS:** Auf ähnliche Weise könnten Unternehmen entsprechende Cloud-Compute-Instanzen mit GPUs und Storage zusammen mit einer kommerziellen oder Open-Source-KI-Plattform bereitstellen. Diese Methode bietet eine ähnliche Kontrolle über die Plattform mit mehr Agilität und einer einfachen Integration in vorhandene Tools. Diese Methode ist möglicherweise die kosteneffizienteste für kleine Bereitstellungen und solche mit unvorhersehbaren oder saisonalen Anforderungen.
- **LLM-API-Services:** Etablierte Services wie OpenAI GPT können verwendet werden, um Funktionen schnell bereitzustellen, ohne dass eine Infrastruktur oder eine KI-Plattform gemanagt werden muss. Diese Methode eignet sich möglicherweise am besten für die Erkundung und den Einstieg, für kleinere Bereitstellungen und solche, die kein hohes Maß an Anpassung oder Steuerung erfordern.

Bevor sich Unternehmen für eine LLM-Plattform entscheiden, sollten sie Zeit investieren, um ihre Anforderungen und benötigten Funktionen zu ermitteln, und einige der folgenden Überlegungen rund um die Auswahl einer Plattform für die LLM-Inferenzierung in Betracht ziehen:

- **Kosten/ROI:** Unternehmen müssen die Kosten und Vorteile der Implementierung und Nutzung jeder Technologieinvestition berücksichtigen. Laut einer Forschungsstudie der Enterprise Strategy Group sind Kosteneinsparungen und ROI die gängigsten Kennzahlen, die Unternehmen nach eigenen Angaben verwenden, um die Effektivität ihrer KI-Initiativen zu messen.[3]
- **Performance und Skalierbarkeit:** Die Dimensionierung der Infrastruktur mit ausreichenden Ressourcen in Prozessoren, GPUs, Arbeitsspeicher und Storage ist wichtig, um sicherzustellen, dass sie die erwartete Parallelität der Inferenzierung bei normalen und Spitzenlasten verarbeiten kann. Außerdem muss die durchschnittliche Inferenzlatenz niedrig genug sein, um NutzerInnen eine positive Erfahrung zu bieten. Unternehmen müssen auch ermitteln, ob das Compute-intensive LLM-Training auf derselben Plattform oder auf einer leistungsstärkeren dedizierten Trainingsplattform durchgeführt wird, bevor auf die Inferenzierungsplattform umgestellt wird.
- **Einfaches Management:** Beim Vergleich einer On-Premise-Infrastruktur mit Cloud-Infrastruktur und -Services ist es wichtig, dass ein Unternehmen seine internen Fähigkeiten berücksichtigt und die Kosten für den Betrieb der Infrastruktur und Plattformen (z. B. Administration, Support und Wartung sowie Stromversorgung/Kühlung) versteht. Mit Colocation-Optionen können Unternehmen viele der Vorteile eines Hostings in den eigenen Rechenzentren nutzen und gleichzeitig die Ressourcen und Kompetenzen auslagern, die für den Betrieb der Infrastruktur und der Plattform erforderlich sind.
- **Erwartete Nutzer-Workloads:** Unternehmen müssen verstehen und prognostizieren, wie viele NutzerInnen auf das Tool zugreifen und wie oft pro Tag sie Fragen stellen werden. Diese wichtige Kennzahl muss bei der Auswahl einer Lösung berücksichtigt werden. Wenn die Nachfrage gering ist, kann ein API-Service ausreichen, aber wenn ein Unternehmen mehr NutzerInnen und Inferenzen unterstützt, ist der Aufbau einer proprietären Plattform kosteneffizienter. Wichtig ist, dass Unternehmen das erwartete Wachstum der Akzeptanz und Nutzungshäufigkeit im Laufe der Zeit berücksichtigen, um sicherzustellen, dass die Infrastruktur angemessen dimensioniert ist und mit den Anforderungen des Unternehmens wachsen kann.

[3] Quelle: Enterprise Strategy Group-Untersuchungsbericht *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

- **Data Governance:** Unternehmen müssen die Standort- und Data-Governance-Anforderungen der Datenquellen berücksichtigen, die zum Trainieren und Pflegen des Modells erforderlich sind. Eine Hybrid-Cloud-Infrastruktur funktioniert am besten, wenn die Daten lokal vorhanden sind oder mühelos dort abgerufen werden können, wo sie benötigt werden. Die Public Cloud kann in einigen Fällen die Erfassung und Zentralisierung von Daten erleichtern. On-Premise-Instanzen ermöglichen es Unternehmen außerdem, die Sicherheit besser zu kontrollieren und die Compliance sensibler Daten sicherzustellen. Das Training mit und die Pflege von Daten, die aktuell, umfassend und unvoreingenommen sind, führt zu einem besseren LLM und genaueren Erkenntnissen, die aus der Inferenzierung abgeleitet werden.

Wirtschaftliche Analyse der Enterprise Strategy Group

Die Enterprise Strategy Group hat eine wirtschaftliche Analyse erstellt, in der die erwarteten Kosten der Inferenzierungsbereitstellung für mehrere Open-Source-LLMs mit RAG und unterschiedlicher Komplexität (einschließlich 7 Mrd. und 70 Mrd. Parametern) sowie für Unternehmen verschiedener Größe (zwischen 5.000 und 50.000 NutzerInnen) verglichen wurden. Dabei wurde davon ausgegangen, dass das Modell eine interne textbasierte Fragen-und-Antworten-Runde bereitstellt und die Inferenzierung dort stattfindet, wo sich die Daten befinden, sodass keine hohen Kosten für eine Datenmigration anfallen. In der Analyse wurden alle Kosten im Zusammenhang mit der Ausführung und Inferenzierung der Modelle über einen Zeitraum von 3 Jahren untersucht, einschließlich der Bereitstellung und des Betriebs der Infrastruktur, der Verwaltung der Systeme und der Zahlung für Cloud-Services, falls erforderlich.

Dell Technologies On-Premise-Infrastruktur im Vergleich zu Public-Cloud-IaaS

Bei unseren Modellen wurden zunächst die erwarteten Kosten für die Ausführung der LLM-Inferenzierung in einer herkömmlichen Infrastruktur (On-Premise-Lösung, in Colocation-Umgebungen, an Edge-Standorten usw.) mit denen der Ausführung in einer ähnlich konfigurierten Public Cloud IaaS auf Amazon EC2-Instanzen verglichen. Die Anforderungen an die Inferenz-Node-Server- und NVIDIA H100-GPU-Konfigurationen wurden für jede Workload basierend auf den Ergebnissen von Inferenz-Baseline-Tests dimensioniert. Damit sollte sichergestellt werden, dass sie die Parallelitätsanforderungen bei regulärer und Spitzenlast (basierend auf den maximalen Anforderungen und der Anzahl der Modellinstanzen) erfüllen und eine angemessene Latenz sowie einen angemessenen Durchsatz für jede erwartete Workload bereitstellen können. Anschließend haben wir alle in Tabelle 1 beschriebenen Kosten sowohl für die Dell Technologies Infrastruktur als auch für die entsprechende EC2-Konfiguration modelliert.

Tabelle 1: Modellierte Kosten und Annahmen für jede LLM-Inferenz-Workload-Anforderung

Kostenkategorie	Dell Technologies (On-Premise-Umgebung)	Public-Cloud-IaaS (Amazon EC2)
Anschaffungskosten (Hardware und Software)	Von Dell Technologies angegebener Preis für Dell PowerEdge R760xa und R660 mit ProDeploy und ProSupport	–
Zusätzliche Kapitalkosten (Zinsen) und Abschreibung (Vorteil)	Im Modell einkalkuliert (8 % WACC, 6 % jährlicher Abschreibungsvorteil)	–
Kosten für Strom und Kühlung	Berechnet auf der Basis von Systemspezifikationen (0,173 USD/kWh)	–
Monatliche Cloud-Ausgaben	–	p5.48xlarge EC2-Instanzkosten, berechnet auf der Basis von Reservierungsrabatten für 3 Jahre
NVIDIA AI Enterprise-Lizenz/-GPU	Basierend auf einer 5-Jahres-Lizenz (anteilig)	Pro Instanz/Std., basierend auf 16 Std./Tag, 5 Tage pro Woche, um die Kosten zu begrenzen
Infrastruktur-/Instanzadministration	Modelliert (10 % bis 100 % der SystemadministratorInnen basierend auf der Anzahl der Nodes)	66 % niedriger als bei einem On-Premise-Modell
ML-Modell- und -Plattformadministration	Modelliert (20 % bis 100 % der ML-TechnikerInnen basierend auf der Anzahl der Modellinstanzen)	Identisch mit On-Premise-Modell

Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Kleineres Modell: Mistral 7B-LLM mit 7 Mrd. Parametern

Für den ersten Vergleich haben wir die Kosten für die Bereitstellung eines kleineren Modells mit etwa 7 Milliarden Parametern modelliert, ähnlich dem Open-Source-LLM Mistral 7B. Zur Dimensionierung der Anforderungen haben wir ein Dimensionierungstool verwendet, das auf den Testergebnissen basierte und die Server- und GPU-Konfigurationen prognostizierte, die eine durchschnittliche Latenz pro Anfrage von etwa 0,4 Sekunden und einen geschätzten Durchsatz von 2,29 bis 6,86 Inferenzen pro Sekunde bereitstellen können. Die allgemeinen Annahmen für die Instanz- und die GPU-Anzahl sind in Tabelle 2 dargestellt.

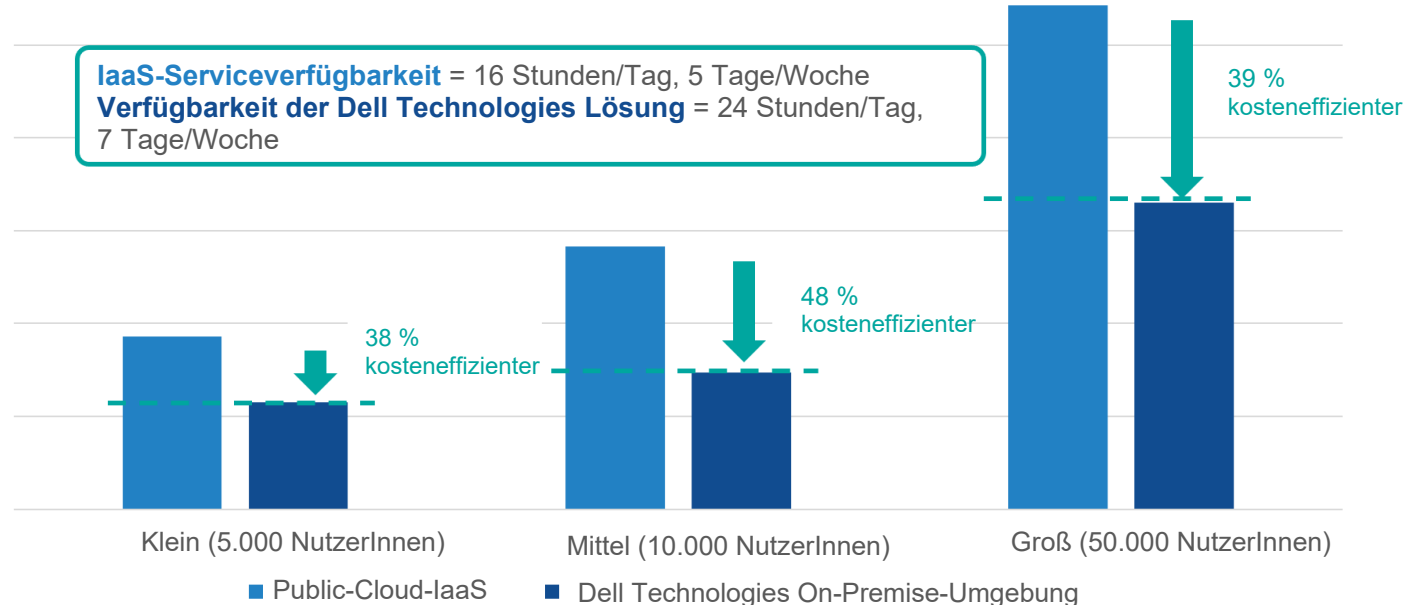
Tabelle 2: Konfigurationsannahmen für die Mistral 7B-Modellinferenzierung mit 7 Mrd. Parametern

LLM-Modell (Anzahl der Parameter)	Anzahl der NutzerInnen	Anzahl der Inferenz-Nodes/-Instanzen	Anzahl der H100-GPUs
Mistral (7B)	5.000	1	1
	10.000	1	2
	50.000	1	4

Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Anschließend haben wir alle in Tabelle 1 zusammengefassten Kosten für jede Konfiguration modelliert. Wie in Abbildung 3 gezeigt, war die Infrastruktur von Dell Technologies 1,6- bis 1,9-mal (38 % bis 48 %) kosteneffizienter bei der Bereitstellung der Inferenzierung für das Unternehmen. Gleichzeitig stand sie dem Unternehmen rund um die Uhr zur Verfügung.

Abbildung 2. Erwartete Kosten für die Inferenzierungsbereitstellung für ein Mistral-LLM mit 7 Mrd. Parametern und RAG



Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Größeres Modell: Llama 2-LLM mit 70 Mrd. Parametern

Anschließend haben wir die zu erwartenden Kosten für die Bereitstellung eines größeren Modells mit 70 Milliarden Parametern modelliert, ähnlich dem Open-Source-70B-LLM Llama 2. Wir haben die Anforderungen erneut mit demselben Dimensionierungstool dimensioniert, um Server- und GPU-Konfigurationen zu prognostizieren, die eine etwas höhere durchschnittliche Latenz pro Anfrage von etwa 1,8 Sekunden und einen geschätzten Durchsatz von 2,29 bis 22,86 Inferenzen pro Sekunde bereitstellen können. Die allgemeinen Annahmen für die Instanz- und die GPU-Anzahl sind in Tabelle 3 dargestellt.

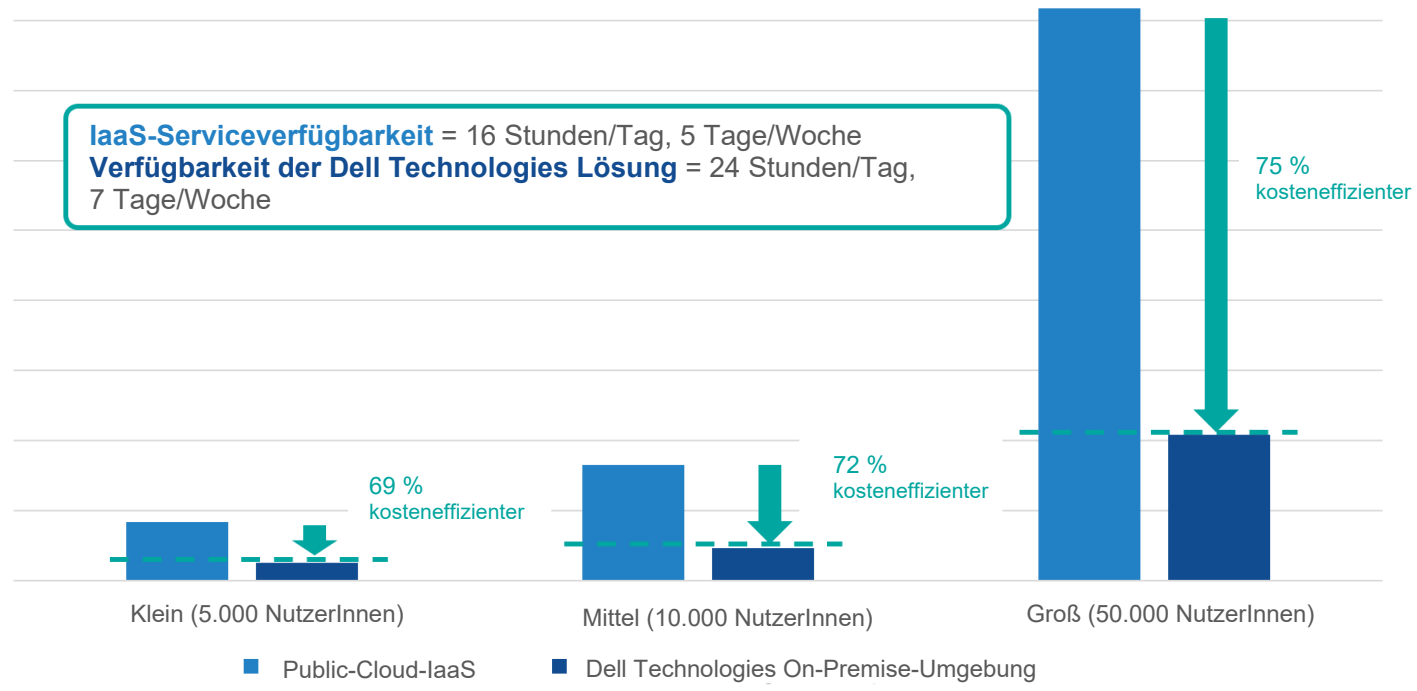
Tabelle 3: Konfigurationsannahmen für die Llama 2-Modellinferenzierung mit 70 Mrd. Parametern

LLM-Modell (Anzahl der Parameter)	Anzahl der NutzerInnen	Anzahl der Inferenz-Nodes/-Instanzen	Anzahl der H100-GPUs
Llama 2 (70B)	5.000	2	8
	10.000	4	16
	50.000	20	80

Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Nach der erneuten Modellierung aller in Tabelle 1 zusammengefassten Kosten für jede der oben dargestellten Konfigurationen stellten wir fest, dass die Infrastruktur von Dell Technologies 3,3- bis 4-mal (69 % bis 75 %) kosteneffizienter bei der Bereitstellung der Inferenzierung für das Unternehmen war. Gleichzeitig stand sie dem Unternehmen rund um die Uhr zur Verfügung.

Abbildung 3. Erwartete Kosten für die Inferenzierungsbereitstellung für ein Llama 2-LLM mit 70 Mrd. Parametern und RAG

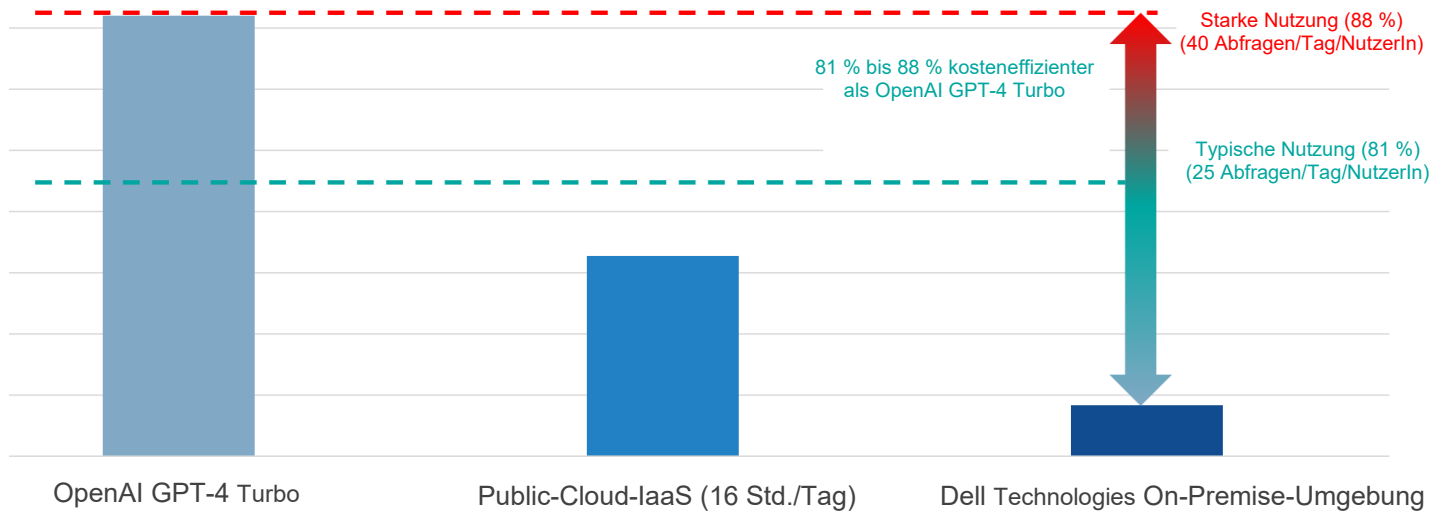


Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Dell Technologies On-Premise-Infrastruktur im Vergleich zum API-basierten GenAI-Service

Anschließend haben wir die erwarteten Kosten für ein großes Unternehmen verglichen, das seinen 50.000 NutzerInnen ein entsprechendes Modell mit 70 Mrd. Parametern bereitstellen möchte. Dafür sollte der etablierte OpenAI-API-basierte GenAI-Service GPT-4 Turbo verwendet werden, der kosteneffizient pro Eingabe- und Ausgabe-„Token“ berechnet wird. Textbasierte Fragen und Antworten erfordern eine moderate Tokenintensität pro Abfrage, weisen keine großen Abweichungen in der Spitzenlast auf und führen zu einem relativ gleichmäßigen Verhältnis zwischen der Anzahl der erforderlichen Eingabe- und Ausgabetoken. Wir sind von insgesamt 1.500 Token (Eingabe plus Ausgabe) pro Abfrage mit durchschnittlich etwa 25 Abfragen pro Tag und NutzerIn sowohl für die On-Premise- als auch für die API-basierte Lösung ausgegangen. Basierend auf unseren Untersuchungen öffentlicher Aussagen haben wir festgestellt, dass dies eine moderate Anzahl von Anfragen pro NutzerIn ist, wobei weniger etablierte Unternehmen weniger Anfragen pro NutzerIn und besser etablierte Unternehmen durchschnittlich bis zu 40 Abfragen pro NutzerIn und Tag erzeugen. Unsere GPT-4 Turbo-Berechnungen haben Kosten von etwa 12,50 US-Dollar pro NutzerIn und Monat prognostiziert, was im Vergleich zu den Suite-basierten KI-Assistenztools, die etwa 30 US-Dollar pro NutzerIn und Monat kosten können, günstig ist. Mit diesen Annahmen haben wir festgestellt, dass die On-Premise-Infrastruktur von Dell Technologies die Inferenzierung 5,4- bis 8,6-mal (81 % bis 88 %) kosteneffizienter als ein API-basierter Service und GenAI-Funktionen für nur etwa 2,31 US-Dollar pro NutzerIn und Monat bereitstellen kann.

Abbildung 4. Erwartete Kosten für die Inferenzierungsbereitstellung für ein Llama 2-LLM mit 70 Mrd. Parametern für bis zu 50.000 NutzerInnen



Quelle: Enterprise Strategy Group, ein Unternehmensbereich von TechTarget, Inc.

Überlegungen

Obwohl die Modelle der Enterprise Strategy Group nach bestem Wissen und Gewissen auf konservativen, glaubwürdigen und validierten Annahmen basieren, wird kein einzelnes modelliertes Szenario jemals jede potenzielle Umgebung repräsentieren. Die Einsparungen für Kunden hängen von ihrem speziellen Anwendungsfall, der Art ihrer Daten, ihrem Fachwissen sowie ihren Modell- und Infrastrukturanforderungen ab. Die Enterprise Strategy Group empfiehlt, dass Sie Ihre eigene Analyse der verfügbaren Produkte durchführen und sich mit Dell Technologies beraten, um die Unterschiede zwischen den Lösungen zu verstehen und zu besprechen, die sich in Ihren eigenen Machbarkeitsstudien bewährt haben.

Dell Technologies für LLM-Inferenzierung

Mit Dell Technologies können Unternehmen auf einfache Weise KI für ihre Daten nutzen, unabhängig davon, wo sie sich befinden. Das bedeutet, dass wir das breiteste Portfolio an KI-Services anbieten – vom Desktop über das Rechenzentrum bis hin zur Cloud –, damit Unternehmen ihre Investitionen richtig dimensionieren und Daten nutzen können, um ihre KI-Fabriken aufzubauen und KI-Anwendungsfälle effizient, sicher und nachhaltig zu realisieren. Dazu bietet Dell Zugriff auf ein umfangreiches Serviceportfolio und ein umfassendes, offenes Ökosystem mit Partnern, die Unternehmen unterstützen, ganz gleich, wo sie sich auf ihrem Weg zur KI befinden – ob sie KI-Strategien erst entwickeln oder ihre GenAI-Investitionen beschleunigen und skalieren möchten.

Für Unternehmen, die mit Datensicherheitsbedrohungen, Compliancebedenken, Datensilos und nicht validierten Datenvolumen konfrontiert sind, können Dell Professional Services für generative KI dazu beitragen, einen Konsens zwischen Geschäfts- und IT-Führungskräften in Bezug auf priorisierte Anwendungsfälle zu schaffen, eine umsetzbare Roadmap zum Erreichen von Zielen bereitzustellen, Unternehmensdaten für die LLM-Integration vorzubereiten, den Reifegrad der Cybersicherheit zu erhöhen und eine KI-Plattform einzurichten, die auf spezifische Geschäftsanforderungen abgestimmt ist. Darüber hinaus können Unternehmen mit Dell APEX KI-Lösungen abonnieren und für Multi-Cloud-Anwendungsfälle optimieren.

Weitere Informationen zu den Lösungen von Dell finden Sie auf der KI-Webseite von Dell.

Fazit

Die erweiterte Nutzung von GenAI in nahezu allen Bereichen des Unternehmens ist ein entscheidender Faktor für einen verbesserten Betrieb und zukünftigen Erfolg. Untersuchungen der Enterprise Strategy Group haben ergeben, dass Unternehmen GenAI derzeit vor allem in den Bereichen Forschung, Marketing, Softwareentwicklung, Produktentwicklung und IT-Betrieb anwenden, und es wird erwartet, dass das Nutzungspotenzial in allen Bereichen steigen wird.[4] Unternehmen können wirkungsvollere und aussagekräftigere Ergebnisse erzielen, indem sie ihre eigene angepasste Version eines LLM trainieren und inferenzieren.

Es gibt mehrere Bereitstellungsmethoden, die für die Inferenzierung von LLMs verwendet werden können, und jede bietet Vorteile für bestimmte Anwendungsfälle und Anforderungen. Für größere Unternehmen mit Tausenden von NutzerInnen, die bereit sind, die Funktionen eines kundenspezifischen LLM zu nutzen, kann die Infrastruktur von Dell Technologies eine leistungsstarke LLM-Inferenzierung bis zu 4-mal kosteneffizienter als IaaS und bis zu 8-mal kosteneffizienter als OpenAI GPT-4 Turbo bereitstellen. Die Enterprise Strategy Group empfiehlt Unternehmen, die LLMs zur Unterstützung ihrer Abteilungen implementieren, die kosteneffizienten Technologien und sachkundigen Services von Dell Technologies zu nutzen, um ein erfolgreiches Ergebnis sicherzustellen, ihre GenAI-Initiativen zu beschleunigen und die Zeit bis zum Erreichen der erwarteten Einsparungen zu verkürzen.

[4] Quelle: Enterprise Strategy Group-Untersuchungsbericht *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

Informationen zu Enterprise Strategy Group
Enterprise Strategy Group von TechTarget bietet fokussierte und umsetzbare Marktinformationen, nachfrageorientierte Studien, Beratungsdienste durch AnalystInnen, GTM-Strategieberatung, Lösungsvalidierungen und kundenspezifische Inhalte, die den Kauf und Verkauf von Unternehmenstechnologie unterstützen.

contact@esg-global.com
www.esg-global.com