

SOUVERÄN GEMACHT

Souverän gemacht #3: Billiger pro Token, teurer im Monat — die drei Stufen der KI-Kostenarchitektur

Von Thomas Faß · 2. September 2026



In dreizehn Tagen haben drei Anbieter Kostenkontrolle in ihre KI-Produkte eingebaut, und die Preislogik selbst hat sich verschoben. Gleichzeitig steigen die Rechnungen weiter und die Ertragswirkung steht still. Diese Ausgabe erklärt, warum beides zusammengehört, und liefert die drei Stufen der Kostenarchitektur sowie die Rechnung, die vor jedem Piloten stehen sollte.

Dreizehn Tage, drei Anbieter, eine neue Produktkategorie

Am 18. August baut Snowflake Kostensteuerung in sein KI-Gateway. Am 26. August führt Google Cloud Ausgabengrenzen für KI-Agenten ein. Und Ende August verschob sich die Preislogik selbst: Zendesk berechnet nur noch Tickets, die die KI vollständig gelöst hat, Pegasystems nimmt Festpreise pro abgeschlossenem Fall und trägt die Modellkosten selbst.

Was hier entstanden ist, hat inzwischen einen Namen: FinOps for AI. Kostenkontrolle wandert von der Nachsorge ins Produkt.

Das lohnt einen genaueren Blick, weil drei Beobachtungen gleichzeitig zutreffen, die sich zu widersprechen scheinen. Die Stückkosten sinken. Die Rechnung steigt. Und der Ertrag steht still.

Was passiert ist

Snowflake wählt jetzt pro Aufgabe automatisch das Modell nach Qualität und Kosten und passt die Wahl an, wenn sich Preise oder Leistung ändern. Der Gedanke dahinter ist unspektakulär und richtig: Für Routineaufgaben braucht niemand ein Frontier-Modell.

Google verkauft acht Tage später Sparpläne mit bis zu 20 Prozent Nachlass auf Token-Kosten, dazu monatliche Ausgabengrenzen und harte Deckel gegen Überschreitungen. Die begleitende Zahl erklärt, warum: Knapp 500 Cloud-Kunden haben binnen eines Jahres über eine Billion Tokens verarbeitet.

Und dann der jüngste Schritt. Wenn ein Anbieter nur noch nach Ergebnis abrechnet, verschiebt er das Kostenrisiko auf die eigene Seite. Das ist für Kunden im Einstieg attraktiv. Gartner bremst allerdings die Erwartung: Heute nutzen 19 Prozent der Services-Einkäufer ergebnisbasierte Abrechnung, und sie steckt in 13 Prozent der Service-Vereinbarungen. Bis 2031 erwartet Gartner, dass Tech-CEOs weniger als ein Viertel ihrer Services-Verträge ergebnisbasiert abschließen. HP plant seine Outcome-Optionen erst für Mitte bis Ende 2027. Gartner-Analyst Tom Coshow bringt es auf den Punkt: „Der Anstieg ergebnisbasierter Preismodelle ist derzeit mehr Hype als Realität.“ (Analystenaussagen, kein ausgewiesenes Erhebungsdesign.)

Drei Anbieter, drei Antworten auf dieselbe Frage. Die Frage lautet: Wie verhindern wir, dass die Rechnung davonläuft?

Warum die Rechnung trotzdem steigt

Die Antwort steckt in einer Zahl aus Nvidias jüngstem Earnings-Call. Jensen Huang beziffert den Rechenbedarf eines KI-Agenten auf das 15- bis 100-Fache dessen, was ein Mensch verbraucht, der ein Werkzeug bedient.

Ein Agent arbeitet in Schleifen. Er prüft, verwirft, versucht es erneut. Genau das macht ihn nützlich, und genau das kostet. Wer von Chat-Nutzung auf Agenten umstellt, wechselt die Größenordnung.

Dazu kommt die Hardware. Nvidia hat die Preise für KI-Server im Juli um 30 Prozent erhöht und kündigt ab Anfang 2027 weitere 15 Prozent an. Treiber sind Speicherchips und Komponentenengpässe. Wer eine eigene Inferenz-Kapazität kalkuliert – und dazu zählt jedes [souveräne Infrastrukturprojekt in Europa](#) –, sollte steigende Hardwarepreise als Annahme in den Business Case aufnehmen. In der Fußnote sind sie falsch aufgehoben.

Die Stückkosten pro Token fallen also weiter. Die Menge steigt schneller. Das ist schlichte Mechanik. Sie fällt nur erst auf, wenn die erste Rechnung nach dem Pilotbetrieb kommt.

Der dritte Baustein: lokale Verarbeitung

Zwischen Cloud-Inferenz und Modell-Routing hat sich in denselben Wochen eine dritte Ebene geschoben. Apple hat mit dem M6 und dem M5 Ultra Chips vorgestellt, die ausdrücklich auf KI-Workloads ausgelegt sind – höhere Speicherbandbreite, verstärkte Neural Accelerators.

Gartner ordnet das so ein: Lokale Verarbeitung nimmt agentische Dauerschleifen von der Cloud-Rechnung. Genannt werden Disney für Kreativteams und Crédit Agricole für Regulatorik-Prozesse, jeweils im Rahmen einer Cloud-to-Device-Strategie, deren Budgets zunehmend aus Infrastruktur-Töpfen kommen.

Für den deutschen Mittelstand ist diese Ebene doppelt interessant. Sie ist günstiger, und sie hält Daten im Haus. Wo ohnehin über Datenschutz nachgedacht werden muss, fallen die Kostenfrage und die Souveränitätsfrage zum ersten Mal zusammen. Bisher zogen sie in entgegengesetzte Richtungen.

Grundlagen: Die drei Stufen der KI-Kostenarchitektur

Weil die Werkzeuge inzwischen bereitstehen, hier die Systematik dahinter. Sie funktioniert im Solution-Design, im Lastenheft und in der Vorstandsvorlage.

Erste Stufe: Cloud-Inferenz. Für alles, was tatsächlich Frontier-Qualität braucht — komplexe Analysen, anspruchsvolle Textproduktion, Aufgaben ohne klares Muster. Diese Stufe ist die teuerste pro Aufruf und in vielen Häusern die einzige, die überhaupt genutzt wird. Genau darin liegt das Problem: Wer alles über ein Spitzenmodell laufen lässt, zahlt Spitzenpreise auch für Aufgaben, die sie nicht brauchen.

Zweite Stufe: Modell-Routing. Pro Aufgabentyp entscheidet ein Gateway, welches Modell zum Einsatz kommt. Klassifikation, Extraktion, Zusammenfassung, Formatierung — für diese Aufgaben genügen kleinere, spezialisierte Modelle, oft zu einem Bruchteil der Kosten. Entscheidend ist, dass die Regel überprüfbar ist: Wer legt fest, welches Modell wann läuft, und was passiert, wenn ein Anbieter seine Preise ändert?

Dritte Stufe: lokale Verarbeitung. Für Dauerläufer und sensible Daten. Ein Agent, der im Hintergrund permanent Dokumente prüft, erzeugt in der Cloud eine Rechnung, die mit der Laufzeit wächst. Lokal erzeugt er Hardwarekosten, die einmal anfallen. Ab welcher Menge sich das dreht, ist eine Rechenaufgabe — und sie lohnt sich häufiger, als man vermutet.

Alle drei sind Architekturentscheidungen. Sie fallen beim Entwurf. Beim Rechnungseingang sind sie längst gefallen.

Die Rechnung, die vor dem Piloten steht

Davor steht allerdings eine unbequemere Frage, und sie hat mit Technik wenig zu tun: **Was kostet bei Ihnen eine einzelne Aufgabe, bevor Sie sie tausendfach automatisieren?**

In den meisten Häusern gibt es darauf keine Antwort. Bitkom hat im Frühjahr 604 Unternehmen ab 20 Beschäftigten befragt. 41 Prozent nutzen KI, 48 Prozent planen es oder diskutieren darüber. Im Sommer 2025 lag die Nutzung bei 36 Prozent, ein Jahr davor bei 20. Die Zahl der Häuser, die KI einsetzen, hat sich in zwei Jahren verdoppelt. Ein Drittel der KI-nutzenden Unternehmen stellt in derselben Erhebung fest, dass KI zu deutlich höheren Kosten geführt hat als erwartet. Ob diese Häuser ihre Kosten je Vorgang kennen, fragt die Erhebung nicht ab — die McKinsey-Zahlen weiter unten legen eine Antwort nahe.

Vier Größen genügen, um das zu ändern:

1. **Was ein einzelner Durchlauf kostet** — einmal tatsächlich gemessen, im Pilotbetrieb.
2. **Wie oft er im Vollbetrieb laufen wird** — realistisch geschätzt, mit Spitzenlast.
3. **Daraus die Monatskosten.**
4. **Dagegengestellt die eingesparte Arbeitszeit, in Euro.**

Bei Agenten kommt eine fünfte Größe hinzu, die Piloten selten zeigen: Wie oft versucht der Agent eine Aufgabe, bevor er aufgibt? Dieser Faktor multipliziert alles darüber, und er erklärt zugleich Huangs 15- bis

100-fachen Rechenbedarf. In einem Pilotbetrieb mit sauberen Testdaten fällt er kaum auf. Im Echtbetrieb mit unvollständigen Eingaben schon.

Drei Fragen für das nächste Anbietergespräch

- Welches Modell läuft je Aufgabentyp, und wer entscheidet das – Sie oder der Anbieter?
- Was passiert bei Budgetüberschreitung: Stopp, Warnung oder Weiterlauf?
- Wie viele Wiederholungen sind pro Aufgabe eingeplant, und was kosten sie?

An den Antworten merken Sie schnell, ob Ihr Gegenüber die eigene Kostenstruktur kennt.

Ein Zusatz für die Praxis: Kostensteuerung ist keine Ja/Nein-Frage fürs ganze Unternehmen. Sie entscheidet sich pro Anwendungsfall. Ein Agent, der zehnmal am Tag läuft, braucht keine Architekturdiskussion. Einer, der zehntausendmal läuft, entscheidet über das Budget des kommenden Jahres. Die Systematik lohnt sich dort zuerst, wo die Menge groß wird.

Die Zahlen, die in keiner Anbieterpräsentation stehen

McKinsey hat für die diesjährige Ausgabe seiner globalen KI-Studie 1.719 Führungskräfte in 97 Ländern befragt. Zwei Ergebnisse daraus gehören nebeneinandergelegt.

28 Prozent der Unternehmen geben inzwischen mehr als ein Zehntel ihres IT-Budgets für KI aus.

6 Prozent erreichen damit eine EBIT-Wirkung von mindestens fünf Prozent.

Der Anteil derjenigen, die überhaupt einen positiven Ergebnisbeitrag sehen, liegt bei 37 Prozent – und ist gegenüber dem Vorjahr unverändert. Gleichzeitig begrenzt ein Fünftel der Befragten die eigene KI-Nutzung wegen der Betriebskosten.

Die Investitionen steigen. Die Wirkung steht still. Und ein wachsender Teil der Unternehmen bremst sich selbst aus, weil die Rechnung nicht mehr passt.

Der Gedanke zum Mitnehmen

Kostensteuerung gehört in den Entwurf. In den meisten Projekten wird zuerst ein Anwendungsfall gebaut und danach gefragt, was er im Betrieb kostet – zu diesem Zeitpunkt sind die teuren Entscheidungen längst gefallen: Modellwahl, Aufrufhäufigkeit, wie oft ein Agent nachfassen darf, bevor er aufgibt.

Ein Solution-Design, das keine Aussage über die Kosten pro Aufgabe trifft, ist kein fertiges Design. Dass Snowflake, Google und inzwischen auch die Abrechnungsmodelle in dieselbe Richtung laufen, ist die Bestätigung von der Anbieterseite. Die Werkzeuge stehen bereit. Es hilft nur wenig, sie nachträglich über eine Architektur zu legen, die ohne sie entworfen wurde.

Diesen Newsletter schreibt Thomas Faß, Gründer von InnoShore aus Münster. InnoShore stellt Senior-IT-Experten bereit und liefert schlüsselfertige IT-Lösungen für den DACH-Mittelstand – mit Senior-Steuerung in Deutschland und einem Compliance-Baukasten, der pro Mandat konfiguriert wird, nicht von der Stange kommt.

Quellen

- Snowflake Dynamic Model Routing im Cortex AI Gateway: Makenzie Holland, [CIO Dive \(18.08.2026\)](#)
- Google Cloud, Sparpläne und Spend Caps für KI-Agenten: [CIO Dive \(26.08.2026\)](#)
- Outcome-Pricing bei Zendesk, Pegasystems und HP; Gartner-Einordnung und Zitat Tom Coshow: Patrick Thibodeau, [CIO Dive \(31.08.2026\)](#)
- Rechenbedarf je Agent (15- bis 100-fach), Aussage Jensen Huang im Q2-Earnings-Call: [Transkript via Investing.com](#)
- Nvidia-Serverpreise, +30 % im Juli und angekündigte +15 % ab 2027: [Computerwoche \(25.08.2026\)](#)
- Apple M6 und M5 Ultra, On-Device-Inferenz und Gartner-Einordnung: [CIO Dive \(25.08.2026\)](#)
- Budget- und EBIT-Zahlen: [McKinsey / QuantumBlack](#), „The State of AI“ 2026 (1.719 Befragte, 97 Länder)
- KI-Nutzung in Deutschland (41 % nutzen, 48 % planen/diskutieren): [Bitkom](#), „Digitalisierung der Wirtschaft: ... Unternehmen beschäftigen sich mit KI“ (11.03.2026) — 604 Unternehmen ab 20 Beschäftigten, telefonisch KW 2–6/2026, repräsentativ; dort auch die Feststellung, dass KI bei einem Drittel der Nutzenden deutlich höhere Kosten verursacht hat als erwartet
- Vergleichswerte 36 % (2025) und 20 % (2024): [Bitkom](#), „Durchbruch bei Künstlicher Intelligenz“ (15.09.2025)

Frühere Ausgaben: [Souveränität ist eine Bauleistung](#) · [Die KI-Kennzeichnungspflicht ist da — reicht das?](#) · [alle Ausgaben im Archiv](#)

SOUVERÄN GEMACHT

Den Newsletter abonnieren

Lesen Sie jede Ausgabe zu KI-Governance und digitaler Souveränität direkt in Ihrem Feed — über unseren LinkedIn-Newsletter „Souverän gemacht“.

[Newsletter auf LinkedIn abonnieren](#)

Die Anmeldung zum Newsletter direkt hier auf der Seite steht in Kürze zur Verfügung.