

KI-Agenten: Kosten skalieren mit Aufgaben, nicht Nutzern

MANAGEMENT SUMMARY & KERNAUSSAGE

KI-Agenten skalieren nicht pro Nutzer, sondern pro erledigter Aufgabe. Inference-Kosten wachsen mit jedem Agenten-Task; Wartung, Monitoring und Compliance machen über 24 Monate oft das Zwei- bis Vierfache der Baukosten aus. Die Betriebskurve ist der entscheidende Hebel – und sie lässt sich planbar machen.

KONTEXT & RELEVANZ FÜR ENTSCHEIDER

Bei klassischer Software wachsen Lizenzkosten mit der Zahl der Nutzer. Bei KI-Agenten ist das anders: Sobald ein Agent eigenständig Recherchen ausführt, Dokumente prüft oder Prozesse übernimmt, entsteht Rechenlast für jede einzelne Ausführung. Altimeter Capital projiziert einen Anstieg der installierten KI-Rechenleistung von rund 18 Gigawatt 2025 auf rund 115 Gigawatt 2028 – mit dem grössten Sprung von etwa 43 Gigawatt allein im Jahr 2027. Für dein Budget heisst das: Die laufende Kostenkurve folgt den erledigten Aufgaben, nicht der Zahl der eingeloggten Mitarbeitenden.

Inference – also die eigentliche Ausführung eines Modells – macht nach Bloomberg Intelligence (via ClearML) bereits rund zwei Drittel des gesamten AI-Compute aus und wächst weiter. Gleichzeitig ist der Stückpreis pro Inference-Einheit in den letzten zweieinhalb Jahren um 99 Prozent gefallen. Das klingt nach Entlastung, ist aber eine Falle: Durch inference-time reasoning steigt der Token-Verbrauch parabolisch. Du kaufst die Einheit günstiger, aber deine Agenten verbrauchen sehr viel mehr davon. Die Gesamtkosten werden damit zu einer Frage des Arbeitsvolumens.

SCHWEIZER UMSETZUNGSKRITERIEN (REVDSG & IT-ARCHITEKTUR)

- ☑ Vollständige Konformität mit dem neuen Schweizer Datenschutzgesetz (revDSG).
- ☑ Keine Abhängigkeiten: Zero Lock-In auf proprietäre geschlossene Modelle.
- ☑ Schnittstellen zu Schweizer ERPs (Abacus, Bexio, M365) im operativen Betrieb.

Unverbindliche Prozess-Analyse

Lass uns in 20 Minuten prüfen, wo KI-Agenten in deinem Unternehmen sofort messbare Entlastung schaffen.

ki-outsourcing.ch