Unternehmens-KI wirtschaftlich steuern

Sinkende Token-Preise bremsen die Ausgaben nicht

Bei agentenbasierten Anwendungen entscheidet nicht allein das eingesetzte Modell über den Gesamtaufwand. Jon Knisley, Head of AI Enablement and Value beim Technologieanbieter Abbyy, zeigt, warum maschinelle Verarbeitung, menschliche Prüfung und technische Umgebung gemeinsam in die Kostenrechnung gehören.

KI hat Software von einem Lizenzmodell mit Fixkosten in einen verbrauchsabhängig abgerechneten Dienst verwandelt. Jeder Prompt, jeder Datenabruf und jede Verarbeitungsschleife fließen in die Abrechnung ein. So steigen die Ausgaben mit der Nutzung. Der eigentliche Kostensprung ist jedoch auf die Architektur zurückzuführen. Der Wechsel vom Chatbot zum Agenten erhöht den Token-Verbrauch pro Aufgabe um das Zehn- bis Hundertfache. Zudem ermöglichen neue Modelle komplexere Abläufe, die mehr Werkzeuge, Aufrufe und Tokens erfordern.

Ein Kommentar von:
John Knisley
Global Product Marketing/Process AI
ABBYY Development Inc.
www.abbyy.com/de

Aufbereitung der Eingaben

Der meistunterschätzte Hebel liegt in der Aufbereitung der Eingaben und der Orchestrierung. Kürzere Prompts, Zwischenspeicherung und die Wahl des Modells bringen begrenzte Einsparungen. Ein großer Teil der Kosten entsteht jedoch durch unnötig generierte Tokens. Bei agentenbasierten Anwendungen entfallen oft nur 5 bis 15 Prozent der verbrauchten Tokens auf die eigentliche Ausgabe. Der Rest ist kontextbedingter Mehraufwand.

Der größte versteckte Kostenfaktor entsteht, wenn Unternehmen die Token-Rechnung statt der Kosten eines erfolgreichen Ergebnisses messen. In die Bezugsgröße muss auch die Arbeitszeit für Spezifikation, Prüfung, Korrektur und Freigabe einfließen. Die Prüfung einer Agentenausgabe durch einen Analysten kostet pro Stunde häufig mehr als der Agent, der sie erstellt hat. Wer Token-Kosten senkt und zugleich den Prüfaufwand erhöht, spart daher kein Geld. Die Ausgaben verlagern sich lediglich auf eine teurere Position.

Technische Optimierung bietet den größeren Hebel

Ein Modellwechsel kann die KI-Kosten senken. Allerdings ist der Effekt geringer, als viele erwarten. In einer Untersuchung blieben die Aufgaben konstant; ausschließlich die technische Umgebung rund um das Modell wurde ausgetauscht. Der Token-Verbrauch sank bei vergleichbarer Qualität um 38 Prozent. Die Kosten pro Aufgabe sanken um 41 Prozent, die Ausführungszeit um 44 Prozent. Der Wechsel vom teuersten zum günstigsten Modell sparte dagegen rund 36 Prozent. Die Optimierung des Modellaufrufs kann somit die Kosten stärker senken als ein Modellwechsel.

Zugleich werden günstigere und Open-Source-Modelle zunehmend praxistauglich. Das gilt besonders, wenn sie saubere, strukturierte Eingaben erhalten, die Mehrdeutigkeiten beseitigen. Nachhaltig ist es, die Wirtschaftlichkeit nicht von einem einzelnen Modell abhängig zu machen. Offene Standards auf Dokumentenebene ermöglichen einen Modellwechsel, ohne dass dafür ein Neuaufbau erforderlich ist.

AI Value Management zur Kostenkontrolle

AI Value Management bedeutet, die Kosten je erfolgreiches Ergebnis zu messen, statt den Token-Verbrauch zu minimieren. Die Berechnung berücksichtigt sowohl die maschinelle Verarbeitung als auch den erforderlichen menschlichen Aufwand. Unternehmen sollten Token-Ausgaben wie ein Portfolio steuern statt als reine Verbrauchskosten zu budgetieren.

Dafür müssen sie die Ausgaben danach bewerten, welchen Wert sie schaffen. Tokens zum Aufbau wiederverwendbarer Fähigkeiten entsprechen Investitionsausgaben. Tokens für interne Abläufe zählen zu den Betriebsausgaben. Tokens in einem kundenorientierten Produkt gehören zu den Umsatzkosten. Sind jedem Prozess ein Verantwortlicher und ein Ergebnis zugeordnet, wird sichtbar, wo KI ihren Wert vervielfacht und wo lediglich Kosten versickern.

Strukturierte Dokumente senken den Verarbeitungsaufwand

Am günstigsten ist ein Token, der gar nicht erst übertragen wird. Unaufbereitete PDF-Dateien zwingen ein Modell, zunächst die Struktur abzuleiten, bevor es die Inhalte auswerten kann. Dadurch steigen Token-Verbrauch, Latenz und Halluzinationsrisiko gleichzeitig. Wird die Datei vorab in ein KI-natives Format überführt, sind Struktur und Bedeutung bereits enthalten. Das Modell kann dann Schlussfolgerungen ziehen, statt das Dokument zu rekonstruieren.

Abbyy-Benchmarks mit dem KI-nativen Format DocLang zeigen je nach Modell eine Senkung der Token-Kosten um den Faktor vier bis über dreißig. Gerade bessere Eingaben ermöglichen es einem kleineren und günstigeren Modell, mit einem Frontier-Modell gleichzuziehen. Die strukturierte Dokumentenaufbereitung bietet dabei den höchsten ROI, weil sie deterministisch erfolgt. Bereits im Dokument enthaltene Fakten lassen sich nahezu ohne Token-Kosten extrahieren. Das teure Reasoning-Modell bleibt dadurch der anspruchsvollen Beurteilung vorbehalten.

Kosten je Ergebnis entscheiden über die Skalierung

Die Branche optimiert weiterhin den Token-Preis, obwohl die Rechnungen im Produktivbetrieb steigen. Forscher sprechen hier von „Tokenmaxxing“. Unternehmen erkaufen zusätzliche Leistungsfähigkeit durch einen höheren Token-Einsatz. Dadurch wächst der Verbrauch pro Aufgabe schneller als deren Wert. Sinkende Preise pro Token verdecken das Problem, bis es bei einer Skalierung sichtbar wird. Deshalb erwartet Gartner, dass bis Ende 2027 mehr als 40 Prozent agentenbasierter KI-Projekte wegen steigender Kosten, unklarem Nutzen und unzureichenden Kontrollen ausgesetzt werden.

Besonders erfolgreiche Unternehmen strukturieren Dokumente und extrahieren deterministische Fakten nahezu kostenlos. Nur echte Beurteilungsaufgaben gehen an ein Sprachmodell. Sie mieten das Modell, verfügen aber über eine eigene technische Umgebung für dessen Einsatz. ◄