02 / RECHENBEISPIEL

Vom Aufruf zur Monatsrechnung

Ein Assistent prüft 100’000 Dokumente pro Monat. Im Beispiel braucht er je Dokument einen GPT-6.1-Sol-Aufruf im Standardtarif. Die durchschnittlichen Werte stammen aus einem fiktiven Nutzungsprofil. [2–5]

Abrechnungsposten	Tokens je Aufruf	USD je Mio.	USD je Aufruf
Reguläre Eingabe	500	2,00	0,00100
Cache lesen	10’000	0,10	0,00100
Cache schreiben, 30 Min.	1’500	2,50	0,00375
Ausgabe inkl. Reasoning	1’500	10,00	0,01500
Modellkosten je Aufruf			0,02075

Monatliche Modellkosten. 100’000 × USD 0,02075 ergeben USD 2’075. Würden alle 12’000 Eingabetokens je Aufruf zum regulären Tarif und die Ausgabe unverändert abgerechnet, ergäben sich USD 3’900. Der Unterschied beträgt USD 1’825 pro Monat. Toolkosten und Steuern kommen je nach Nutzung hinzu.

USD 20,75 je 1’000 Dokumente. Das sind die Modellkosten im dargestellten Nutzungsprofil. Integration, Betrieb und menschliche Prüfung sind darin nicht enthalten.

Das Beispiel setzt voraus, dass grosse Teile der Eingabe wiederverwendet werden können. Bei wechselnden Dokumentinhalten oder seltenen Aufrufen kann die Ersparnis deutlich kleiner ausfallen. Übernehmen Sie die Cache-Anteile deshalb aus einem gemessenen Pilotbetrieb.

So gleichen Sie die Rechnung ab

Lassen Sie sich die Verbrauchsdaten getrennt nach regulärer Eingabe, Cache-Lesen, Cache-Schreiben und Ausgabe bereitstellen. Multiplizieren Sie jede Menge mit dem jeweiligen Tarif und teilen Sie bei Preisen je Million durch 1’000’000. Addieren Sie anschliessend allfällige Zusatzgebühren. Die gemeldeten Ausgabetokens enthalten die Reasoning-Tokens bereits. [2, 3]

Wann der Cache spart

Setzen Sie unveränderte Anweisungen oder Referenztexte an den Anfang, wechselnde Kundendaten ans Ende. Für die gezeigten Modelle braucht es mindestens 1’024 sichtbare Eingabetokens am Anfang. Der tiefere Tarif greift bei einem Treffer. Wird der Eintrag kaum wiederverwendet, können Schreibkosten die Einsparung aufzehren. [3]