Tokens

Die Einheit, in der ein Sprachmodell rechnet und abrechnet

Ein Sprachmodell versteht keine Wörter, es rechnet mit Zahlen. Dafür zerlegt es jeden Text in Tokens und gibt jedem eine feste Nummer. Diese Einheit bestimmt danach fast alles: wie viel in eine Anfrage passt, wie schnell die Antwort kommt und was sie kostet.

VOM SATZ ZUR ZAHLENFOLGE

1. **Zerlegen.** Der Text wird in Bausteine geschnitten. Ein Token kann ein ganzes Wort sein, eine Silbe oder ein einzelnes Zeichen. Auch Satzzeichen zählen mit.
2. **Nummerieren.** Jeder Baustein bekommt eine feste Nummer. Erst damit kann das Modell rechnen. Es sagt dann die wahrscheinlichste nächste Nummer voraus.
3. **Bedeutung zuordnen.** Jede Nummer wird in eine Zahlenreihe übersetzt, die ähnliche Wörter nah beieinander anordnet. Erst hier entsteht so etwas wie Bedeutung.

EIN BEISPIEL

ACHT TOKEN

„Guten Morgen, wie kann ich helfen?“ wird zu:
Guten | Morgen | , | wie | kann | ich | helfen | ?

Lange Wörter zerfallen in Teile:
„Weiterbildungsangebot“ wird zu Weiter | bildungs | angebot.

DAS KONTEXTFENSTER IST EINE HARTE GRENZE

WAS HINEINMUSS
Der Prompt, die Systemvorgaben im Hintergrund, der bisherige Gesprächsverlauf und die Antwort selbst. Alles zusammen.

WAS PASSIERT
Liegt die Grenze bei 8.000 Token und verbraucht der Prompt schon 6.000, bleibt für die Antwort fast nichts übrig.

WAS ZU TUN IST
Lange Dokumente in Abschnitte teilen. In Dialogen den Verlauf regelmäßig zusammenfassen oder kürzen.

WARUM DAS AUF DER RECHNUNG LANDET

Größe	Wie sie wirkt	Was daraus folgt
Abrechnung	Meist je 1.000 oder je 1.000.000 Token, Eingabe und Ausgabe getrennt gezählt	Eine lange Antwort kostet genauso wie eine lange Frage
Modellgröße	Ein großes Modell kostet ein Vielfaches pro Token	Für Zusammenfassen und Einordnen reicht oft das kleine
Menge	Bei automatisierten Massenabfragen summieren sich Centbeträge	Vor dem Rollout hochrechnen, nicht danach
Rate Limits	Anbieter begrenzen Token je Minute oder Tag	Die Architektur muss Lastspitzen aushalten, etwa zum Monatsabschluss

FÜNF HEBEL GEGEN UNNÖTIGEN VERBRAUCH

1. **Früh messen.** Den Token-Zähler des Anbieters auf typische Prompts anwenden und interne Richtwerte festlegen, bevor etwas produktiv geht.
2. **Prompts straffen.** „Ich möchte, dass du mir hilfst, einen Text zu schreiben, der sehr professionell klingt und sich an KMU richtet“ wird zu „Schreibe einen professionellen Text für die Zielgruppe KMU.“
3. **Kontext filtern statt fluten.** Erst die passenden Textstellen suchen, dann nur diese mitgeben. Das senkt den Verbrauch und macht die Antwort genauer.
4. **Das passende Modell wählen.** Zusammenfassen und Einordnen können auch kleine Modelle. Ein günstiges kann vorschalten und nur die schweren Fälle weiterreichen.
5. **Im Team Regeln setzen.** Wer verbraucht wofür wie viel? Volltexte vor der Eingabe kürzen.

MERKSÄTZE

Ein Token ist nicht dasselbe wie ein Wort. Und die Antwort kostet extra, sie wird meist getrennt abgerechnet.

Mehr Kontext ist nicht besser. Gefilterter Kontext ist besser.

Bei Massenabfragen ist der Tokenverbrauch keine technische Nebensache, sondern eine betriebswirtschaftliche Größe.