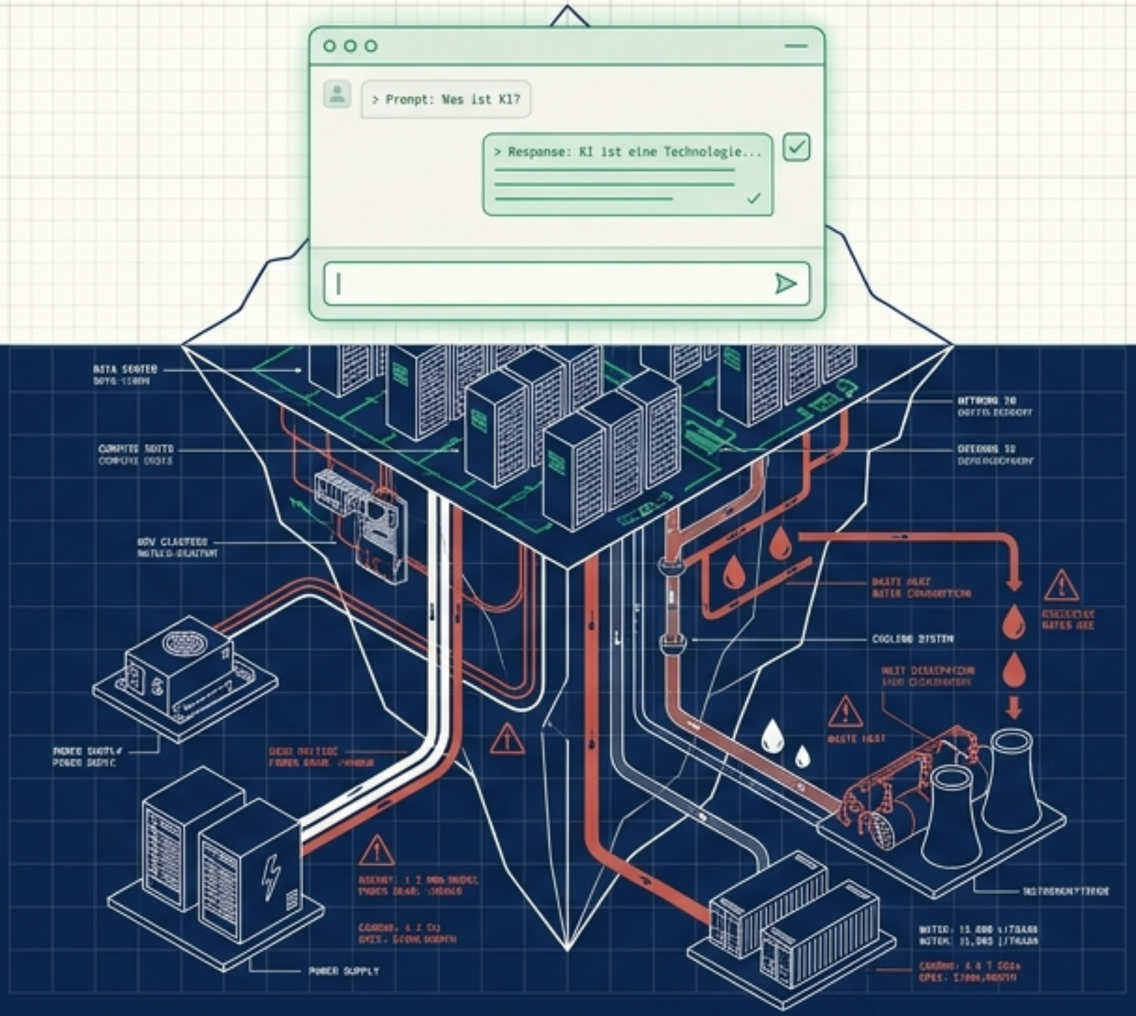
TOKENS · KOSTEN · RESSOURCEN

Was eine KI wirklich verbraucht – und was es überhaupt braucht, um sie zu nutzen.



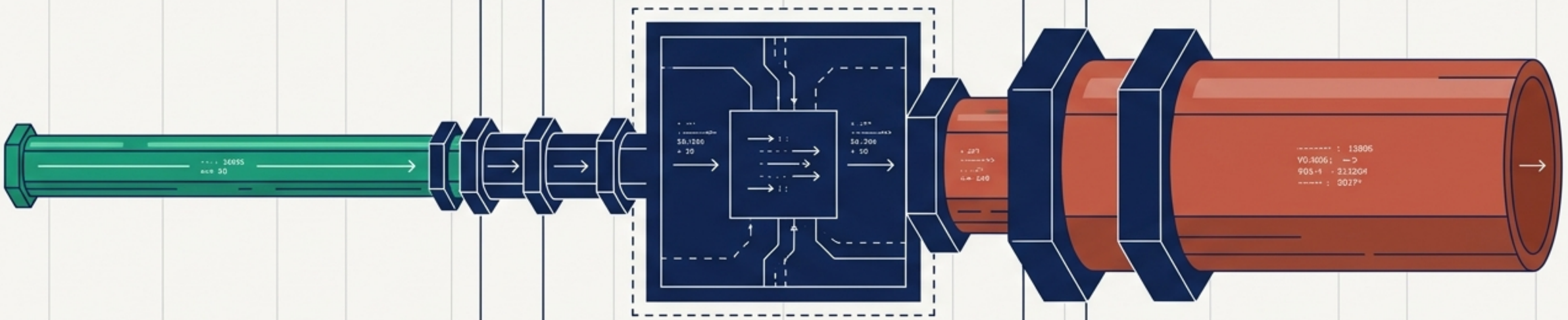
Erstellt: Claudia Heil mit Claude

Die unsichtbare Währung der KI.

Input-Tokens(Eingabe)

Alles, was hineingeht: Ihre Frage, der Kontext, der Chatverlauf und vor allem hochgeladene Dateien.

Interne Verarbeitung



Nur bei Denkmodellen (Reasoning): Eine unsichtbare Gedankenkette, die mitberechnet und mitbezahlt wird.

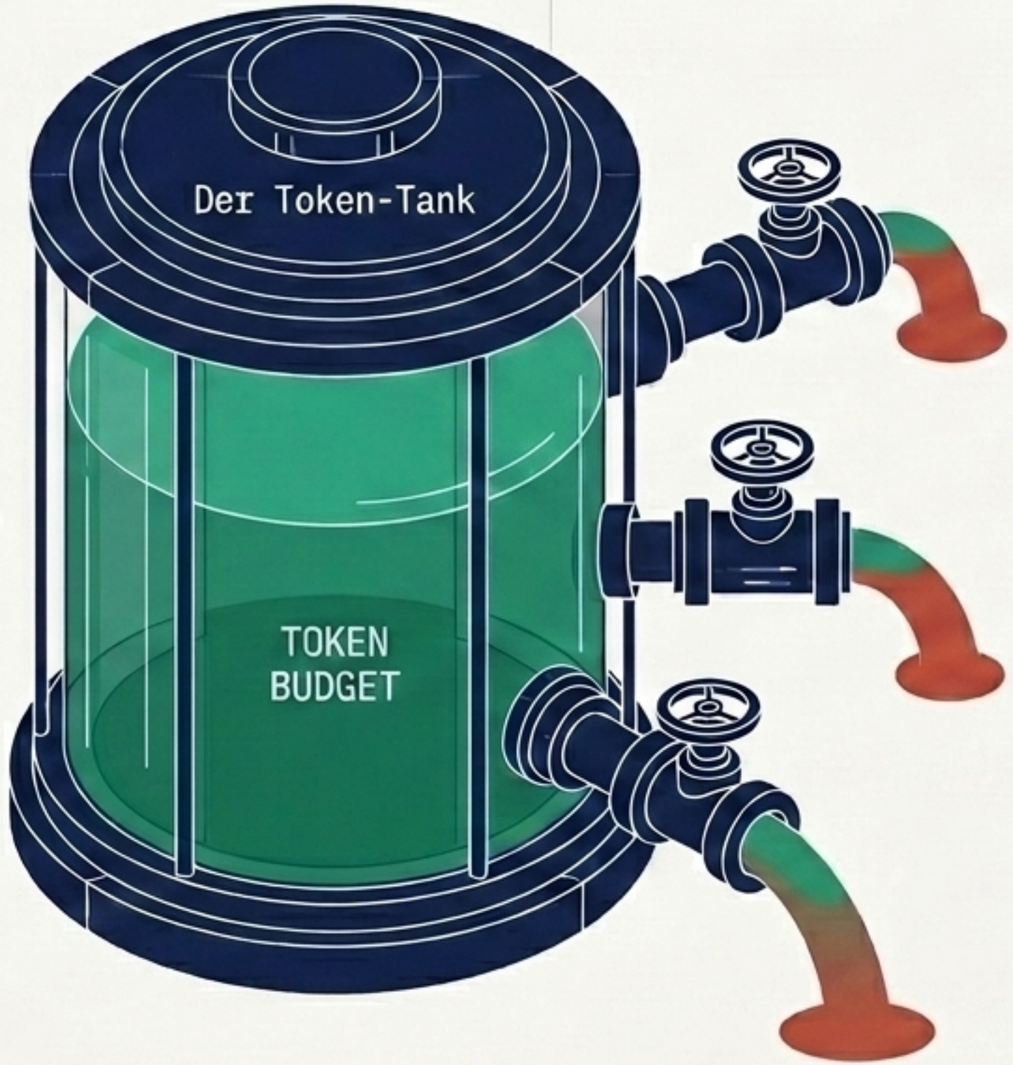
Output-Tokens (Ausgabe)

Alles, was die KI zurückschreibt. Achtung: Output ist meist 3- bis 5-mal teurer als Input.

Merksatz: Sie zahlen für Eingabe und Ausgabe – und bei Denkmodellen für das stille Mitdenken dazwischen.

Wo das Token-Budget unbemerkt versickert.

Drei alltägliche Gewohnheiten leeren den Tank besonders schnell.



Große Anhänge

Ein 20-Seiten-PDF = ca. 13.000–18.000 Tokens – und das bei jeder Nachricht erneut.

Lange Antworten

Output ist der teure Teil. Länge präzise vorgeben heißt Kosten steuern.

Interationen

Jede Nachfrage schickt den gesamten bisherigen Verlauf erneut mit. Anhänge zählen jedes Mal wieder mit.

Die richtige Hebelwirkung für jede Aufgabe.

KLEINE MODELLE GPT-4o mini, Claude Haiku, Gemini Flash, Llama 8B	GROSSE MODELLE GPT-4o/5, Claude Opus, Gemini Pro, Llama 405B
Schnell, sehr günstig, energiesparend.	Tiefe Analyse und gutes Urteilsvermögen.
5- bis 30-mal günstiger als große Modelle.	Stark bei Unsicherheit und Mehrdeutigkeit.
Für Routine: umformulieren, sortieren, zusammenfassen, Format ändern.	Für Analyse, Strategie, schwierige Texte, kniffligen Code.

Faustregel: Klein für Routine und Struktur – groß für Analyse und Unsicherheit.

Ein Markt im permanenten Wandel.

Preise fallen rasant.

In 18 Monaten um Faktor 280 gefallen (Quelle: Stanford AI Index 2025).

Modelle wechseln.

Neue Generationen, Namen, Gratis-Kontingente und Tarifgrenzen ändern sich laufend.

Achtung.

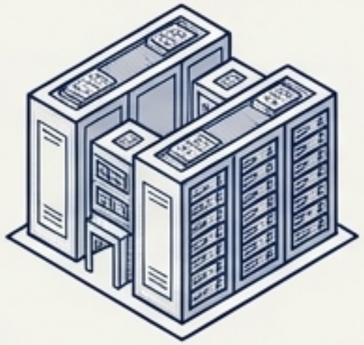
Preise, Limits und Kontingente ändern sich in Monaten – nicht Jahren.

Konsequenz: Zahlen direkt vor der Nutzung prüfen; Prüfdatum in Projekten stets notieren.

Die massive Maschinerie hinter einem leeren Textfeld.

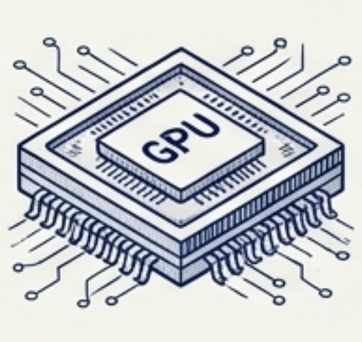
Bevor ein einziges Token verarbeitet wird, muss eine komplette globale Infrastruktur bereitstehen:

Rechenzentren



Fußballfeldgroße Hallen. Ein großes Zentrum verbraucht bis zu 1 GW – wie eine Großstadt mit ~800.000 Haushalten (z.B. München).

Spezial-Chips



Zehntausende GPUs pro Modell. GPT-4 = ~25.000, Llama 3.1 = ~16.000. Ein Cluster = ca. 1 Mrd. $ Hardware.

Strom

Rechenzentren brauchten 2024 ca. 415 TWh – rund 1,5 % des Weltstroms (IEA 2025).

Wasser

Zur Kühlung. Verdunstet zu ~80 % und ist dem lokalen Kreislauf entzogen.

Daten

Training braucht Billionen Tokens. Llama 3.1 nutzte 15 Bio. – das entspricht ~18 Mio. Bibeln an Text.

Der physische Fußabdruck einer einzigen Anfrage.

Highest

Lowest

Video

100–700+ Wh = Mikrowelle 1 Stunde laufen lassen (für einen 5-Sek-Clip). Extrem teuer, wächst rasant.

Bild

ca. 2–3 Wh = 1 Smartphone-Akku voll laden. (20–40× mehr als Text).

Audio

ca. 1–5 Wh = Handy 10–40 % laden. (Sprache erzeugen/erkennen).

Text

ca. 0,3 Wh = LED-Lampe
ca. 2 Min. | ca. 0,3 ml = Rund 5 Tropfen Wasser.
(Sparsamste Anwendung).

20–50 Anfragen = 0.5L Wasserflasche.

Faustregel: Je mehr Pixel und Sekunden, desto teurer – Video doppelt so lang bedeutet rund viermal so viel Verbrauch.

Der reale Ressourcenverbrauch im Modellvergleich.

Modell / Typ	Strom je Anfrage	Wasser je Anfrage	Einordnung
Llama 3.1 8B (klein, offen)	ca. 0,03–0,1 Wh	unter 2 ml	Sehr sparsam – läuft fast auf dem PC
Gemini Text (Google)	ca. 0,2–0,3 Wh	ca. 0,26 ml direkt	Sparsam dank effizienter TPU-Chips
ChatGPT / GPT-4o (OpenAI)	ca. 0,3 Wh	ca. 0,3 ml direkt	Standard, gut gemessen (Altman 2025)
Claude (Anthropic)	ca. 0,4–0,6 Wh (geschätzt)	Keine Angabe	Größenordnung wie ChatGPT
Reasoning-Modelle (o3, GPT-5)	7–40 Wh	20–100 ml	Bis ~100× mehr – Nachdenken kostet immens

Wichtig: Anbieter veröffentlichen kaum offizielle Zahlen – die Werte sind Richtwerte.

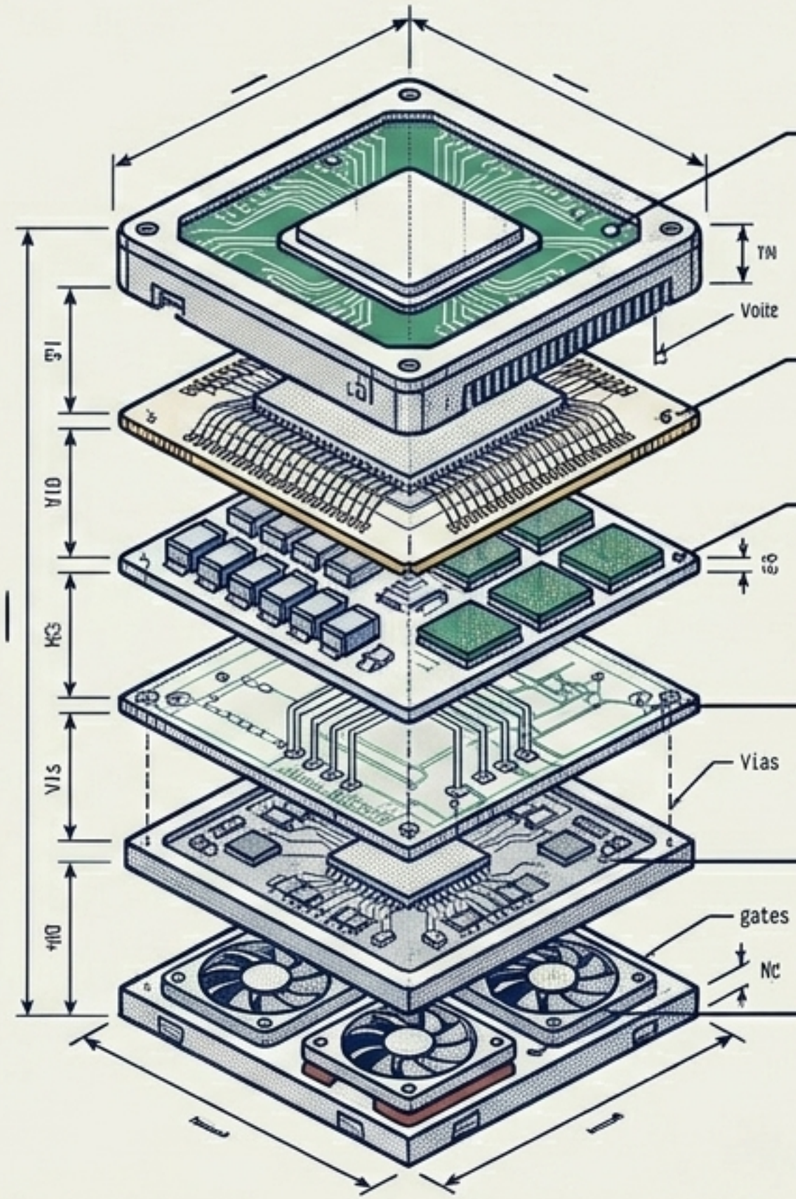
Die Anatomie eines KI-Gehirns.

Was ist eine GPU?
Ein Grafikprozessor, der extrem viele kleine Aufgaben gleichzeitig rechnet – das Herzstück jeder KI.

Base Stats

Preis: Wie ein Mittelklasse-Auto.

Wasser bei Herstellung: 2.000-8.000 Liter je Chip.



- **Kupfer & Aluminium:** Leiterbahnen, Kühlkörper, Gehäuse.
- **Gold & Silber:** Kontakte und feine Verbindungen.
- **Kobalt & Tantal:** Speicher und Kondensatoren (oft Konfliktrohstoffe).
- **Wolfram:** Verbindungen zwischen den Chip-Schichten.
- **Silizium:** Grundmaterial des eigentlichen Chips.
- **Seltene Erden (Neodym):** Magnete in Lüftern und Speichern.

Fünf goldene Regeln für den KI-Alltag

1. Eingabe vs. Ausgabe	Sie zahlen für beides. Output ist 3- bis 5-mal teurer; Denkmodelle berechnen zusätzlich eine unsichtbare Gedankenkette.
2. Den Token-Tank schützen	Anhänge, Länge und Iterationen treiben den Verbrauch. Große Dateien werden bei jeder Nachricht neu berechnet.
3. Das richtige Modell wählen	Klein für Routine, groß nur für Analyse und Unsicherheit. Kleine Modelle sind 5- bis 30-mal sparsamer.
4. Die Skalierung begreifen	Ein einzelner Prompt (Text) kostet nur ~fünf Tropfen Wasser. Aber das Training (z.B. 70.000 Eimer für GPT-3) und Milliarden Nutzer summieren sich extrem.
5. Physische Realität anerkennen	Tokens sind nur die Spitze. Dahinter stehen gigantische Rechenzentren, immenser Strombedarf und Rohstoffe aus aller Welt.