Das Kontextfenster: Arbeitsfläche, nicht unendliches Gedächtnis

Token bestimmen, was die KI sehen, verstehen und verarbeiten kann.

Token verstehen: Kontext, Kosten und Qualität in der KI-Arbeit steuern

Was wir heute vorhaben

BLOCK 1

Einführung in das Thema Token

- Was ist das?
- Was sollte man Wissen?

BLOCK 2

Token-Grundlagen

- Tokenizer & Co.

BLOCK 3

Token, Ressourcen, Modelle & Tools

- Energieverbrauch
- Wasserverbrauch
- Modellauswahl
- Kosten

BLOCK 4

Praxis, Transfer, Checkliste

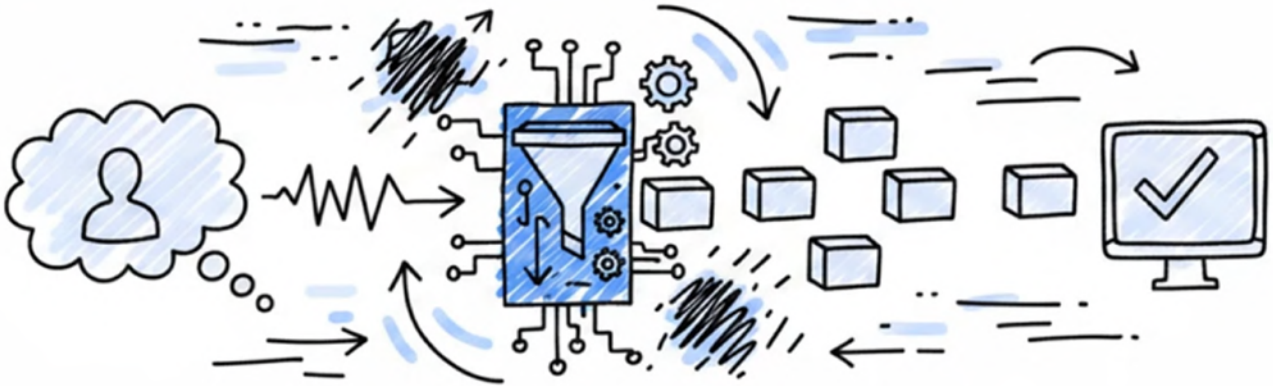
- Best Practices
- Use Cases
- Dies & das

17:00 Einstieg · 20:00 Ende

Token

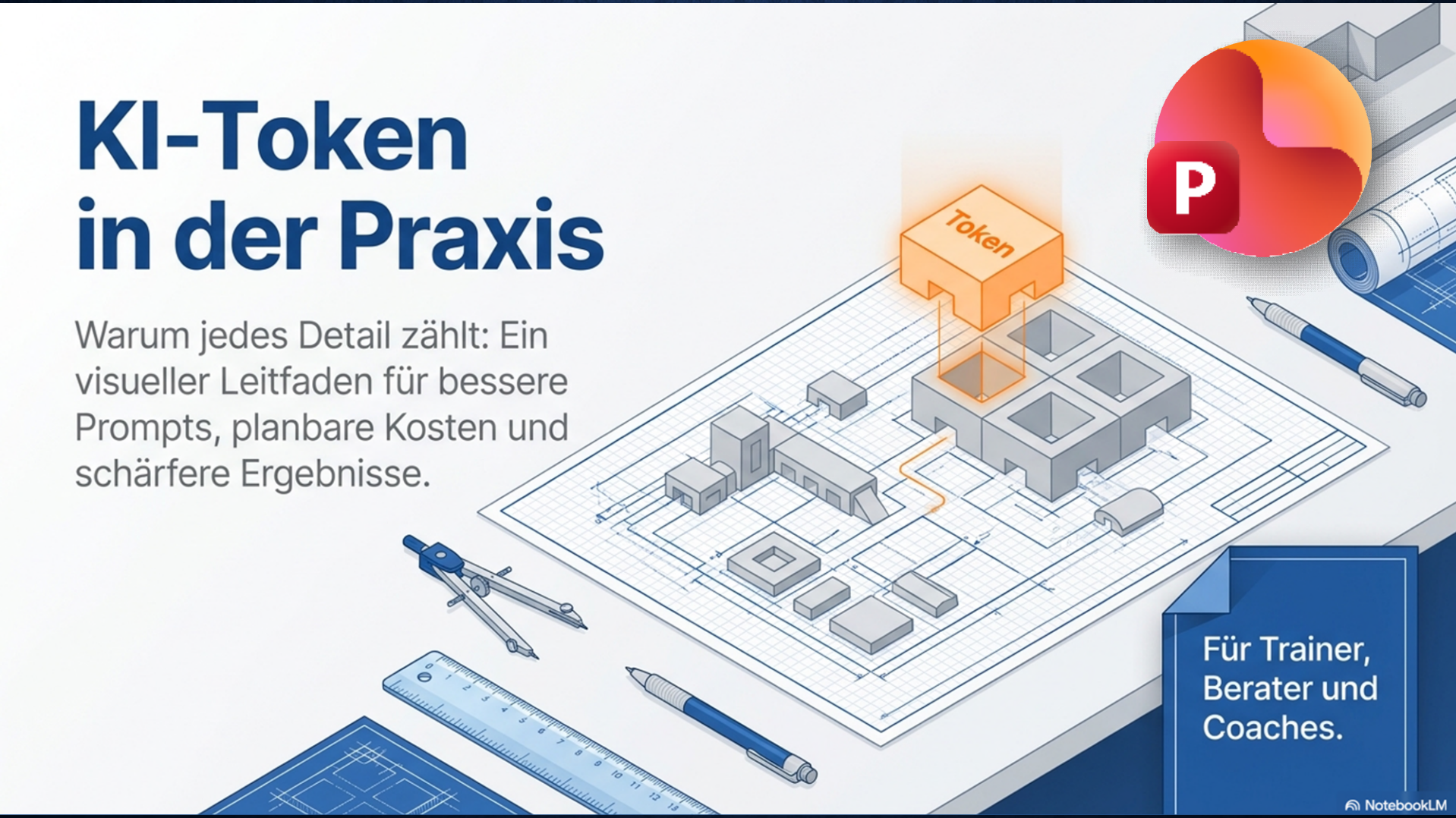
- Einführung in das Thema

KI-Token meistern

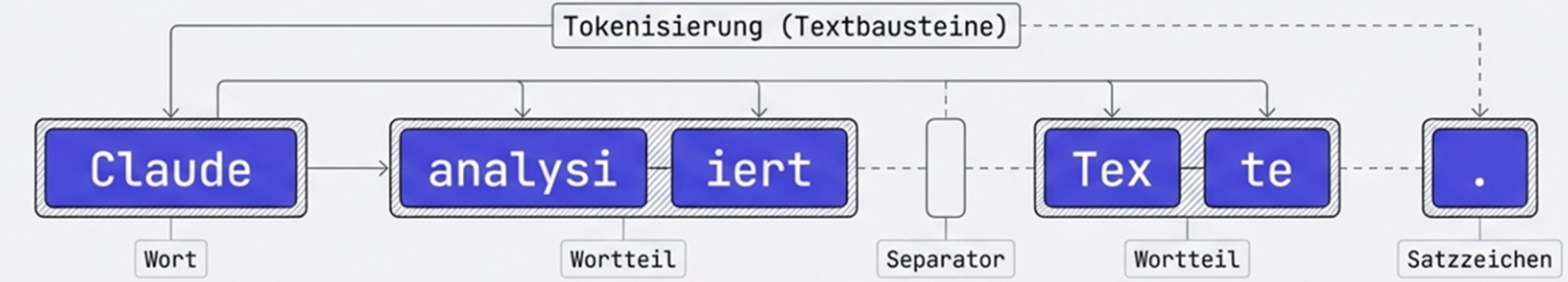


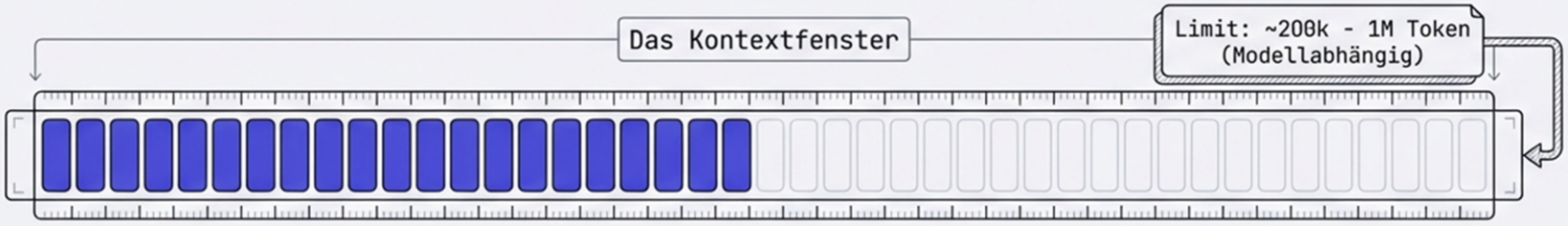
KI-Token in der Praxis

Warum jedes Detail zählt: Ein visueller Leitfaden für bessere Prompts, planbare Kosten und schärfere Ergebnisse.



Die Anatomie des Inputs: Token & Kontext





Key Data Points

- **Tokenisierung:** Die Grundeinheit der Verarbeitung (ca. 0,7 Wörter).
- **Gedächtnis:** Begrenzt auf das aktuelle Fenster (Working Memory).
- **Kosten:** Berechnung erfolgt pro Input- und Output-Token.

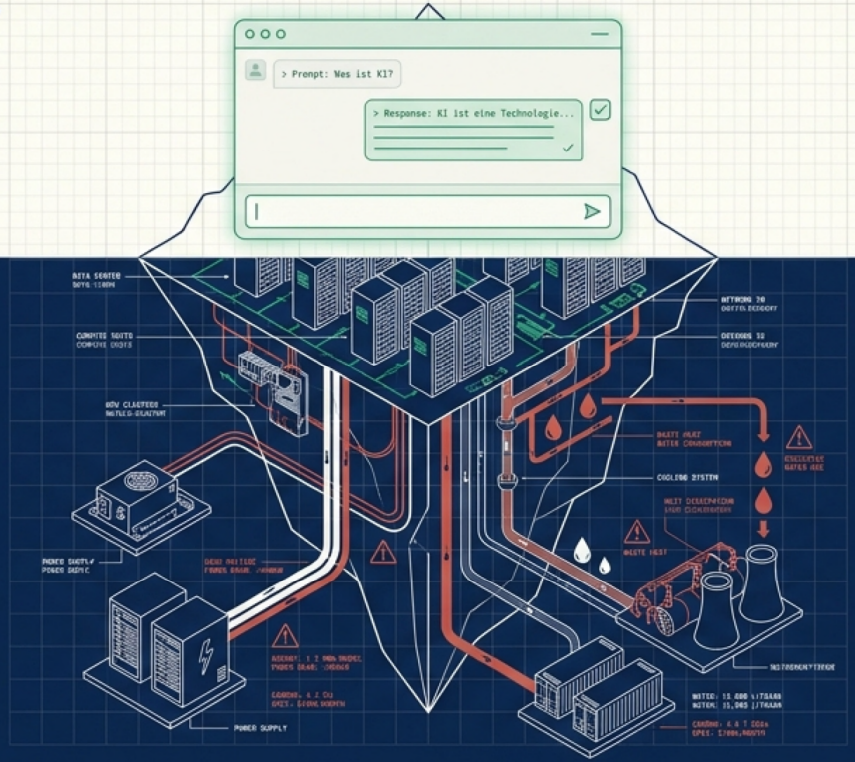
Token

- Grundlagen

- Token
  - Ressourcen
  - Energieverbrauch
  - Kosten
  - Modellwahl

TOKENS · KOSTEN · RESSOURCEN

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



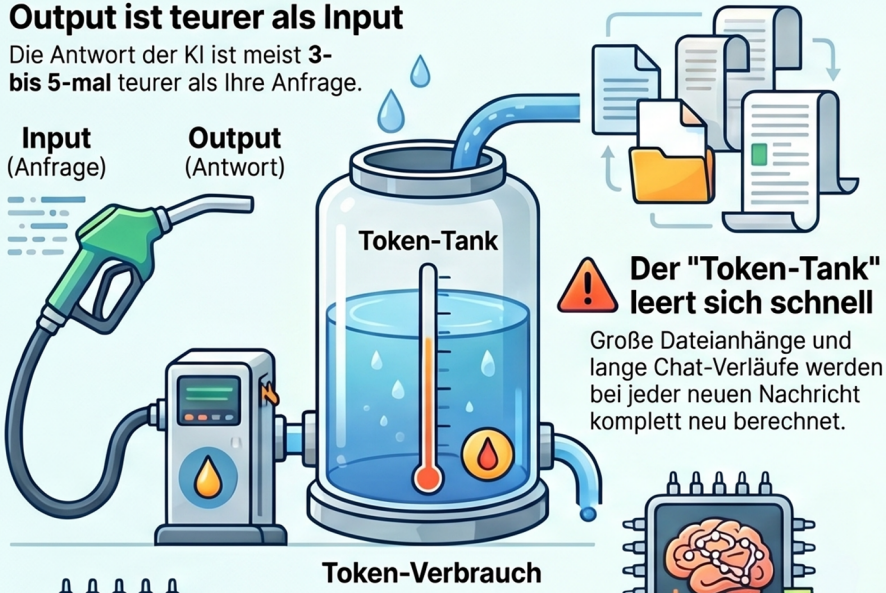
Erstellt: Claudia Heil mit Claude

KI-Effizienz: Kosten, Ressourcen und Modellwahl im Überblick

Kostenlogik & Verbrauchstreiber

Output ist teurer als Input

Die Antwort der KI ist meist **3- bis 5-mal** teurer als Ihre Anfrage.



Der "Token-Tank" leert sich schnell

Große Dateianhänge und lange Chat-Verläufe werden bei jeder neuen Nachricht komplett neu berechnet.

Reasoning-Modelle kosten extra

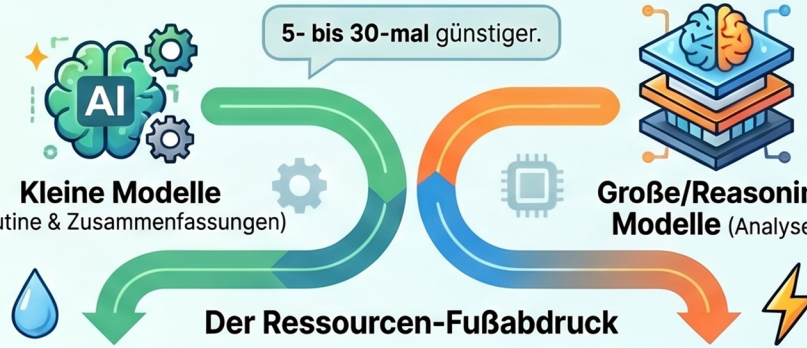
Bei Denkmodellen wird die unsichtbare interne Gedankenkette als zusätzlicher Verbrauch mitbezahlt.

Warnhinweis

Aktualität & Achtung: Preise ändern sich in Monaten, nicht Jahren

Strategische Modellwahl & Ressourcen

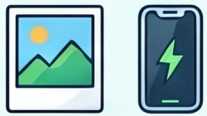
Faustregel: Klein für Routine, groß für Analyse



Text

ca. 0,3 Wh
LED-Lampe (2 Min.)

Bild

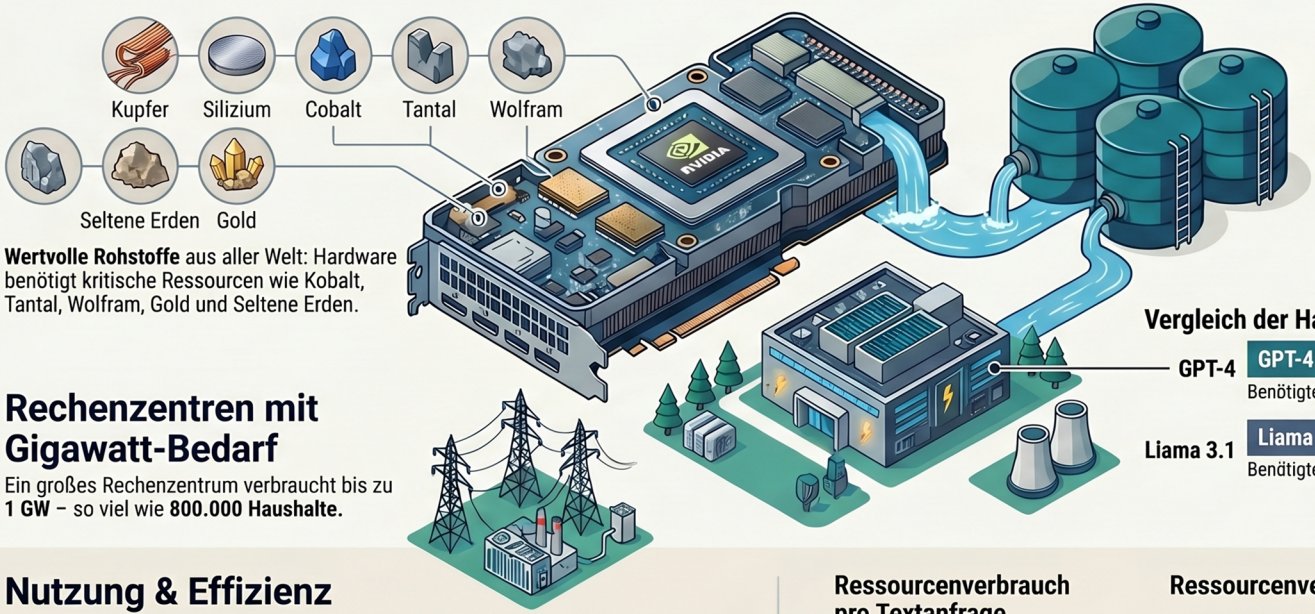


ca. 2–3 Wh
1 Smartphone-Akku laden

Video ist der "Verbrauchs-Riese"

ca. 100–700+ Wh
Mikrowelle (1 Std.)

Der unsichtbare Preis der KI: Ressourcen, Kosten & Hardware



Wertvolle Rohstoffe aus aller Welt: Hardware benötigt kritische Ressourcen wie Kobalt, Tantal, Wolfram, Gold und Seltene Erden.

GPU-Herstellung: Enormer Wasserfußabdruck

Die Produktion eines einzeinen KI-Chips (z. B. NVIDIA H100) verbraucht **2.000–8.000 Liter Wasser.**

Rechenzentren mit Gigawatt-Bedarf

Ein großes Rechenzentrum verbraucht bis zu **1 GW** – so viel wie **800.000 Haushalte.**

Vergleich der Hardware-Anforderungen großer Modelle

GPT-4

GPT-4: ca. 25.000

Benötigte Spezial-Chips (GPUs) | rund **1 Mrd. $** Hardware-Wert (Cluster)

Liama 3.1

Liama 3.1: ca. 16.000

Benötigte Spezial-Chips (GPUs) | rund **1 Mrd. $** Hardware-Wert (Cluster)

Nutzung & Effizienz

Output ist 3- bis 5-mal teurer als Input

Die Antwort der KI kostet deutlich mehr als die gestellte Frage.

Ressourcenverbrauch pro Textanfrage

0,3 ml Wasser / 5 Tropfen

ca. 0,3 Wh Strom / LED-Lampe leuchtet 2 Min.

Ressourcenverbrauch nach Medientyp

Text: ca. 0,3 Wh

Bild: ca. 2–3 Wh
1 Smartphone-Akku laden

Video (5 Sek): 100–700+ Wh
Mikrowelle läuft 1 Std.

Modellwahl als größter Sparhebel: Kleine Modelle sind 5- bis 30-mal günstiger und energiesparender als Analyse-Modelle.

Token

- Kontextfenster
- Arbeitstisch & Fokus


Die Kunst des Prompt-Designs
Wie Sie LLMs wie ein Bildhauer steuern – Präzise, effizient und mit maximalem Fokus.
NotebookLM

Praxis & Transfer

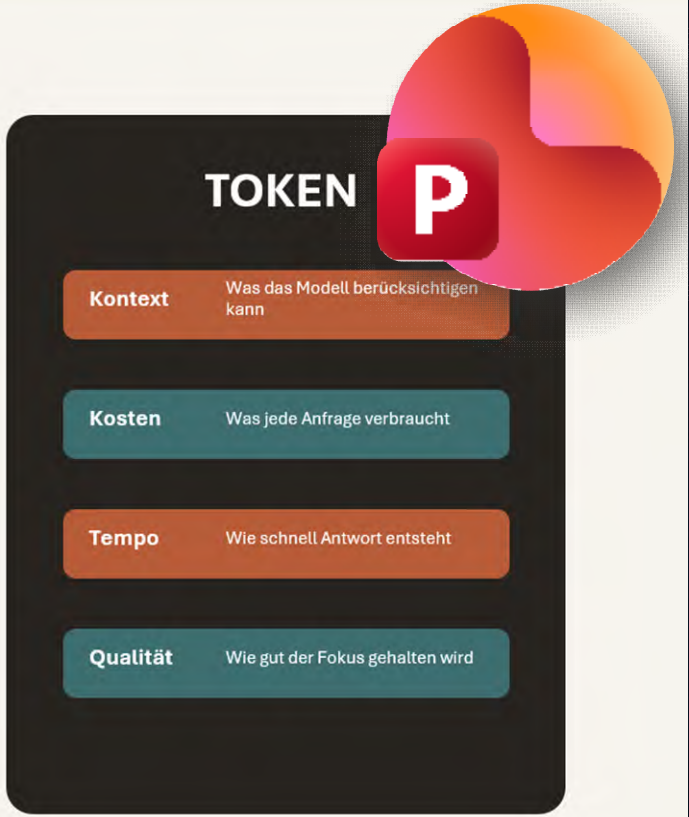
Strukturieren, Kürzen, Absichern

Ein Leitfaden für Token-bewußtes und präzises Arbeiten

Token bewusst einsetzen

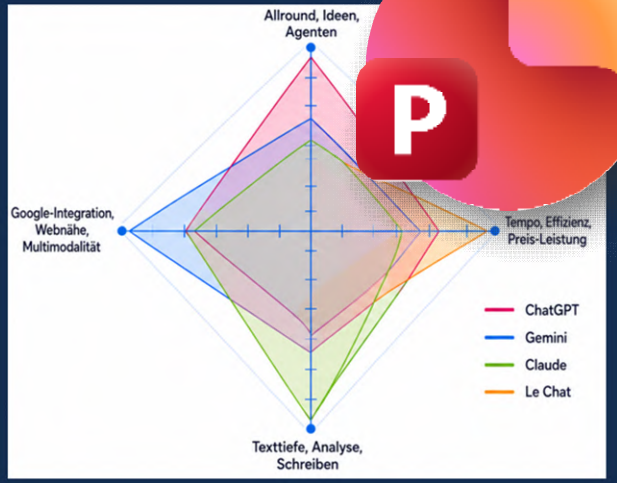
Best Practices für das Arbeiten mit LLMs

Kernidee: nicht maximal füllen, sondern gezielt steuern.



Faustregel: So viele Token wie nötig, so wenige wie möglich.

Entscheidungsmatrix
KI-Tools



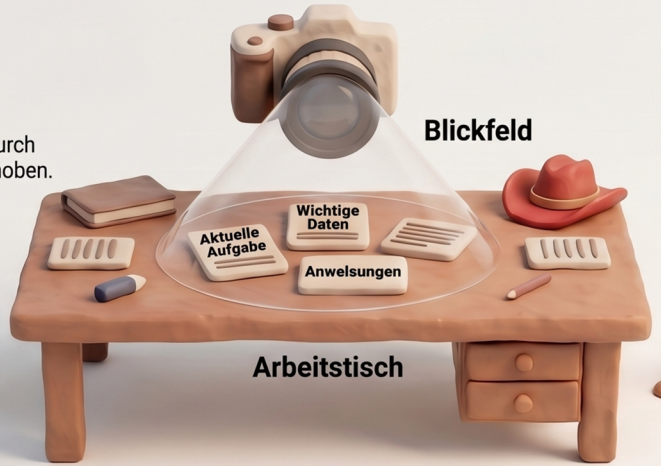
Modellvergleich und Auswahlkriterien

OpenAI · Google Gemini · Anthropic Claude · Mistral / Le Chat

Stand: Mai 2026

Das KI-Kontextfenster: Deinen digitalen Arbeitstisch effizient nutzen

Verdrängung alter Informationen
In langen Verläufen werden frühere Anweisungen durch neue Informationen physisch aus dem Fenster geschoben.



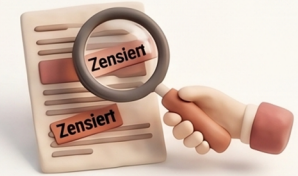
Verwässerung der Prioritäten
Ohne klare Struktur verliert die KI den den Fokus auf das Wichtige.

Strategien für die Praxis

Struktur durch Gliederung
Nutze klare Überschriften und Prioritäten zur Orientierung.

Zwischenfazits ziehen
Verdichte lange Verläufe regelmäßig in kurze Zusammenfassungen.

Relevanz-Check vor dem Upload
Trenne strikt zwischen relevanten Daten und störendem Rauschen.



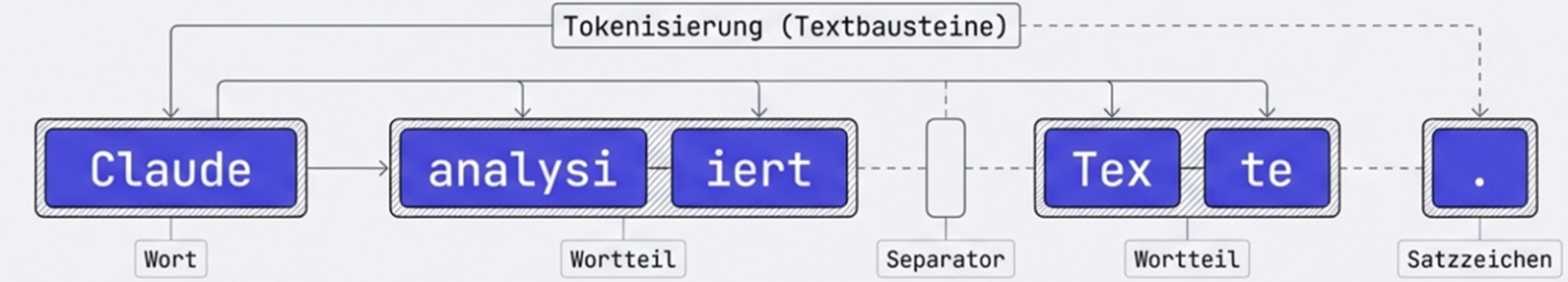
Datenschutz-Regel
Prüfe kritisch, welche Daten notwendig sind und welche draußen bleiben müssen.

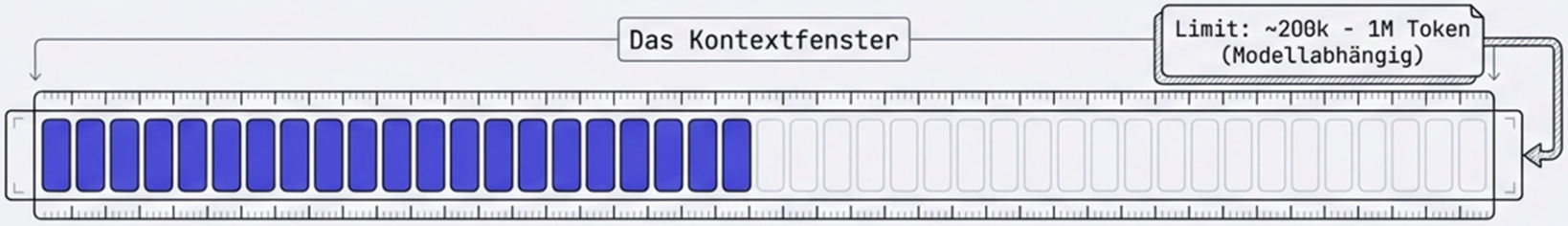
Übungen zur Anwendung

Die Sortierübung
Trainiere das Trennen von relevantem Material und Störmaterial.

Prompt-Umbau
Übe das Komprimieren eines langen Chat-Verlaufs in ein klares Zwischentazit.

Die Anatomie des Inputs: Token & Kontext

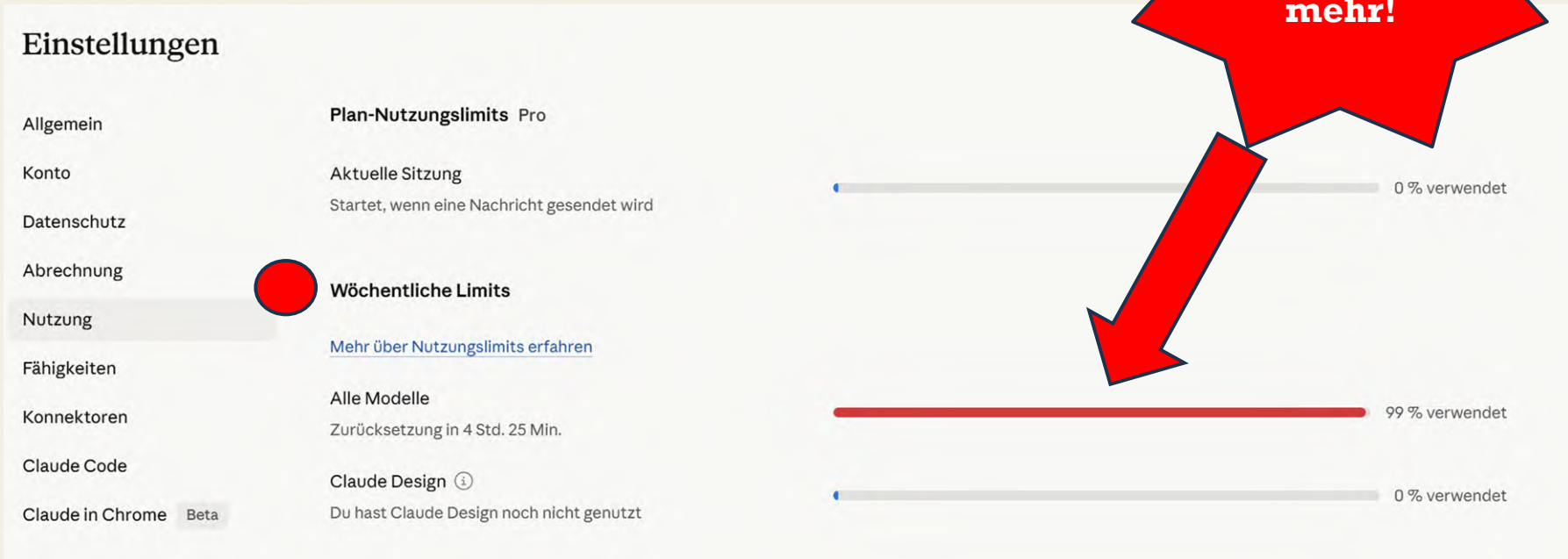




Key Data Points

- **Tokenisierung:** Die Grundeinheit der Verarbeitung (ca. 0,7 Wörter).
- **Gedächtnis:** Begrenzt auf das aktuelle Fenster (Working Memory).
- **Kosten:** Berechnung erfolgt pro Input- und Output-Token.

Auf Token-Verbrauch achten!



KI-Kontext steuern und meistern.

Wie Sie das Gedächtnis und den Fokus von KI-Modellen gezielt lenken.

Ein Praxis-Modul für Trainer, Berater und Coaches.

WORKSHOP-BEGLEITUNG

Arbeiten mit Claude

Token verstehen, klüger arbeiten.

Begleitend zum Workshop
„PowerPoint-Präsentationen mit Claude erstellen“

Bernhard Laukamp · Trainertreffen Deutschland