HANDOUT GRUNDLAGEN TOKEN & KONTEXT

Token & Kontext

Warum ChatGPT vergisst — und was du dagegen tun kannst

5 Abschnitte · Erstellt: März 2026 · Version 1.0

Was sind Tokens?

Die Grundeinheit aller Sprachmodelle

> Sprachmodelle wie ChatGPT lesen keinen Text so wie wir. Sie zerlegen jeden Satz in kleine Einheiten — sogenannte Tokens. Ein Token liegt irgendwo zwischen einer Silbe und einem Wort.

BEISPIEL: TOKENISIERUNG

Das Wort **"Unvorhergesehen"** wird so zerlegt:

Un vor her ge se hen

1 deutsches Kompositum = 6 Tokens

💡 Faustregel: 1 deutsches Wort ≈ **1–3 Tokens**

TOKEN-GRÖSSEN IM ALLTAG

Inhalt	Tokens (ca.)	Kontext
Kurze Frage	~15	Minimaler Verbrauch
Detaillierter Prompt	~80	Normaler Arbeitsauftrag
E-Mail + Anweisung	~500	Standard-Anwendungsfall
150-seitiges PDF	~80.000	Füllt ein Kontextfenster erheblich

💬 **Fun Fact:** Deutsch verbraucht deutlich mehr Tokens als Englisch. "Kraftfahrzeughaftpflichtversicherung" = 8 Tokens. "Car insurance" = 2 Tokens. Im Alltag ist das aber **kein Grund, auf Englisch zu prompten** — der Unterschied ist zu klein, um zu zählen.

> *„ Sprachmodelle verstehen keine Wörter. Sie rechnen mit Tokens. “*

Das Kontextfenster

Der Arbeitsspeicher des Modells — und warum er begrenzt ist

> Das Kontextfenster ist die maximale Menge an Tokens, die ein Modell auf einmal „sehen" kann. Deine Frage, der bisherige Chatverlauf, hochgeladene Dokumente und Systemanweisungen — alles zusammen muss reinpassen. Was nicht reinpasst, existiert für das Modell nicht.

METAPHER: DAS GLAS

Stell dir das Kontextfenster wie ein Wasserglas vor. Alles was du hineinschüttest wird verarbeitet. Aber je mehr du hineinschüttest, desto mehr läuft über. Was überläuft, ist weg — ohne Vorwarnung.

SO WÄCHST DER KONTEXT IN EINEM CHAT

Nachricht 1 — Erste Frage ~10 Tokens
Fast leer. Maximaler Spielraum.
Nach ein paar Nachrichten ~420 Tokens
Das Modell schickt jedes Mal den **gesamten Verlauf** erneut mit.
Langer Chat + PDF-Upload ~120.000 Tokens
Zunehmend ineffizient — Antworten werden langsamer und ungenauer.
Überlauf — Fenster voll Anfang gelöscht
Die ältesten Nachrichten werden still entfernt. Kein Hinweis, keine Warnung.

> „ *Je länger das Gespräch, desto weniger erinnert sich das Modell an den Anfang.* “

Warum ChatGPT vergisst

Der Mechanismus dahinter — einfach erklärt

> ChatGPT hat kein echtes Gedächtnis. Jede Antwort wird komplett neu berechnet — und dabei bekommt das Modell jedes Mal den gesamten bisherigen Chatverlauf mitgeliefert. Das ist kein Bug. Das ist das Design.

WAS WIRKLICH PASSIERT

1. Du stellst eine Frage

Deine Nachricht wird als Tokens kodiert und ins Kontextfenster gelegt.

2. Das Modell berechnet eine Antwort

Es betrachtet alles im Fenster: deine Frage, den bisherigen Verlauf, hochgeladene Dateien.

3. Du stellst die nächste Frage

Jetzt wird **alles erneut mitgeschickt** — Frage 1, Antwort 1, Frage 2, Antwort 2, plus deine neue Frage. Das Fenster wächst mit jeder Runde.

4. Das Fenster ist voll

Die ältesten Nachrichten werden automatisch entfernt. Das Modell hat keine Wahl — und informiert dich nicht darüber.

> *„ Lange Chats machen Antworten nicht besser. Kurze, fokussierte Chats schon. “*

Wichtige Nuance: Das Ziel ist nicht Kürze — sondern Relevanz. Ein langer, fokussierter Chat mit relevantem Kontext ist besser als viele kurze neue Chats ohne Kontext. Was irrelevant ist: raus. Was relevant ist: drin lassen. Kürze ist kein Wert an sich.

Praktische Tipps

4 Strategien — sofort anwendbar, kein Tool nötig

> Du brauchst kein neues Tool und keine Einrichtung. Diese vier Strategien wirken sofort — mit dem, was du heute schon hast.

DIE 4 STRATEGIEN

Strategie 1

Irrelevanten Kontext weglassen

Ein 150-seitiges PDF füllt 80.000 Tokens. Oft reichen die 3 relevanten Seiten. Alles was nicht direkt zur Frage gehört, verdünnt die Antwortqualität.

WANN NICHT? Wenn Hintergrundinfo tatsächlich relevant ist — nicht kürzen. Relevanter langer Kontext ist besser als kein Kontext.

Strategie 2

Neuen Chat starten

Wenn das Thema wechselt oder der Chat lang wird: neuer Chat = frisches Kontextfenster. Sofort bessere Ergebnisse, ohne Konfiguration.

WANN NICHT? Wenn der bisherige Verlauf viel relevanten Kontext enthält — dann besser zusammenfassen (Strategie 3) statt wegwerfen.

Strategie 3

Chat zusammenfassen lassen

Lass das Modell den aktuellen Stand zusammenfassen. Dann neuer Chat mit dieser Zusammenfassung. Gleiche Information — Bruchteil der Tokens.

WANN NICHT? Wenn der Chat noch kurz ist — Aufwand lohnt sich erst nach längeren Gesprächsverläufen mit vielen Entscheidungen.

Strategie 4

Wissensmanagement-Tools

Tools wie NotebookLM halten Dokumente als dauerhafte Wissensbasis — ohne den Chat zu füllen. Der Kontext bleibt frei für deine Fragen.

WANN NICHT? Für Einmalaufgaben oder kurze Recherchen ist der Tool-Setup unverhältnismäßig — direkter Chat reicht vollständig.

PROMPT-VORLAGE: ZUSAMMENFASSUNG FÜR NEUEN CHAT

Fasse unseren bisherigen Gesprächsverlauf in maximal 200 Wörtern zusammen. Fokus auf: getroffene Entscheidungen, offene Punkte, wichtige Fakten. Diese Zusammenfassung nutze ich als Startpunkt für einen neuen Chat.

PROMPT-VORLAGE: DOKUMENT GEZIELT ANALYSIEREN

Ich lade dir Seite [X–Y] aus diesem Dokument hoch. Beantworte nur diese Frage: [konkrete Frage]. Ignoriere alles was nicht direkt relevant ist.

„ Neuer Chat, weniger Ballast, bessere Antworten. So einfach ist Kontextmanagement. “

Wie gut ist mein Ergebnis?

Formal korrekt ≠ wirklich gut — woran du Kontextprobleme erkennst

> Ein langer Chat produziert nicht zwingend schlechte Ergebnisse — aber ein Chat der zuviel irrelevanten Ballast enthält, schon. Der Unterschied liegt nicht in der Länge, sondern in der Relevanz des Inhalts.

VORHER / NACHHER

Formal korrekt

Neuen Chat für jede Folgefrage

Du startest jedes Mal von vorne — kein Kontext, kein Verlauf, keine Entscheidungen aus dem letzten Chat. Das Modell muss alles neu aufbauen.

Wirklich gut

Ein fokussierter Chat mit relevantem Verlauf

Kontext bleibt erhalten — Entscheidungen, Rahmenbedingungen, Einschränkungen. Irrelevantes wurde vorher rausgehalten. Das Modell arbeitet mit dem, was zählt.

DREI SIGNALE FÜR EIN KONTEXTPROBLEM

Das Modell „vergisst" frühere Entscheidungen. Wenn Antworten plötzlich nicht mehr zu dem passen, was vorher besprochen wurde — ist das Kontextfenster zu voll oder zu leer.
Antworten werden generischer. Je mehr irrelevantes Material im Chat ist, desto weniger kann sich das Modell auf das Wesentliche konzentrieren.
Du musst ständig wiederholen. Wenn du dieselben Rahmenbedingungen immer wieder erklärst, fehlt ein strukturierter Kontext-Einstieg.

> *„ Nicht kürzer, nicht länger — relevanter. Das ist das Ziel. “*

Modell-Übersicht

Kontextfenster-Größen der wichtigsten Modelle

> Verschiedene Modelle haben unterschiedlich große Kontextfenster. Je größer das Fenster, desto mehr Text kann das Modell gleichzeitig verarbeiten — ohne zu vergessen.

KONTEXTFENSTER-VERGLEICH (VEREINFACHT)

Modell	Kontextfenster	Entspricht ca.	Ideal für
GPT-4o OpenAI	128k Token	≈ 300 Seiten	Standard-Anwendungen, normale Chats, kurze bis mittlere Dokumente
Claude 3.5 Anthropic	200k Token	≈ 500 Seiten	Große Dokumente, lange Analysen, komplexe Aufgaben
Gemini 1.5 Pro Google	1.000k Token	≈ 2.500 Seiten	Ganze Bücher, sehr große Datensätze, umfangreiche Codebasen

ℹ️ **Achtung:** Ein größeres Kontextfenster bedeutet nicht automatisch bessere Antworten. Modelle können mit vielen Tokens überfordert sein und relevante Details übersehen. Qualität schlägt Quantität.

WANN WELCHES MODELL?

💼 **Alltag & E-Mails**

GPT-4o reicht vollständig. Kein Grund zu wechseln.

Lange Berichte oder Verträge

Claude 3.5 mit 200k Tokens ist hier die bessere Wahl.

Ganze Bücher oder Datenbanken

Gemini 1.5 Pro mit 1 Million Tokens — für Extremfälle.

> *„ Das richtige Modell für die richtige Aufgabe — nicht immer das größte. “*