

RAG ERKLÄRT: WARUM KI RÄT UND WIE DU IHR DIE RICHTIGEN INFOS GIBST

3 PROMPTS

TOOL: CLAUDE + CHATGPT

STAND 2026-10-02

Du verstehst, warum ChatGPT, Claude und Gemini manchmal selbstbewusst falsche Antworten geben, und wie du das mit RAG abstellst: Die KI sucht erst in deinen Dokumenten und antwortet dann. Danach richtest du es in vier Wegen selbst ein, vom einzelnen PDF im Chat bis zum Claude-Projekt mit hunderten Seiten. Das wichtigste Stück zum Kopieren ist der Anti-Rate-Prompt: Er zwingt die KI, die Fundstelle zu zitieren oder offen zu sagen, dass es nicht drinsteht.

4 Wege

VOM PDF BIS ZUR WISSENSBASIS

0 €

EINSTIEG IN ALLEN DREI TOOLS

10 Min

BIS ZUR ERSTEN ANTWORT MIT BELEG

WAS DU BRAUCHST

Ein Konto bei mindestens einem Tool: ChatGPT, Claude oder Google (für NotebookLM). Alle drei haben einen kostenlosen Einstieg.

Ein oder mehrere Dokumente, zu denen du Fragen hast: PDF, Word-Datei, Text oder eine Website. Für den ersten Test reicht eine Bedienungsanleitung oder ein Mietvertrag.

Optional: **Claude Pro** für rund 21 € im Monat (18 € netto plus MwSt.). Brauchst du nur, wenn dein Projekt so groß wird, dass Claude in den RAG-Modus schaltet.

Optional: **ChatGPT Plus** für 23 € im Monat, wenn du mehr als 5 Dateien pro Projekt brauchst.

WARUM KI LIEBER RÄT

Andrej Karpathy, Mitgründer von OpenAI, hat gesagt, dass Kontext-Engineering so ziemlich die wichtigste Fähigkeit im Umgang mit KI ist. Im Original (X, 25. Juni 2025) nennt er es „the delicate art and science of filling the context window with just the right information for the next step“. Auf Deutsch: die Kunst, der KI im richtigen Moment genau die richtigen Infos hinzulegen.

Um zu verstehen, warum das so wichtig ist, musst du wissen, wie ein Sprachmodell arbeitet. Fachwort: **LLM**, kurz für Large Language Model. Ein LLM schreibt Wort für Wort. Bei jedem Schritt rechnet es aus, welches Wort als nächstes am wahrscheinlichsten passt. Es schlägt dabei nichts nach. Alles, was es „weiß“, steckt in Mustern aus dem Training.

Fragst du: „Wo sitzt bei uns der IT-Support?“, rechnet das Modell trotzdem. Vielleicht kommt „im 3. Stock“ mit 62 % raus und „im Keller“ mit 21 %. Die Antwort klingt sicher, ist aber geraten. LLMs geben dir trotzdem eine Antwort, weil sie so gebaut sind, dass sie immer die wahrscheinlichste Antwort geben.

OpenAI hat selbst rausgefunden, dass KI lieber rät, als zuzugeben, dass sie etwas nicht weiß. Im Paper „Why Language Models Hallucinate“ (September 2025) vergleichen die Forscher Sprachmodelle mit Schülern in einer schweren Prüfung: Wer rät, hat eine Chance auf Punkte. Wer das Feld leer lässt, bekommt sicher null. Modelle werden ähnlich bewertet, also lernen sie zu raten. Diese selbstbewussten Falschantworten heißen **Halluzinationen**.

DAS BEISPIEL MIT JULIUS

Stell dir das so vor. Julius weiß ziemlich viel über die Welt und kann dir zum Beispiel sagen, wie groß der Tiergarten ist. Aber wenn du ihn fragst, wo in deiner Firma der IT-Support sitzt, hat er natürlich keine Ahnung.

Der Unterschied zur KI: Julius tut nicht so, als wüsste er es. Er sagt einfach „keine Ahnung“. Das LLM dagegen rät.

Außer du gibst ihm eure Firmendokumente. Dann kann auch Julius darin suchen, die richtige Stelle finden und dir korrekt antworten, statt zu raten. Genau das macht deine KI auch: Sie holt sich erst den passenden Kontext und generiert dann die Antwort.

RAG: RETRIEVAL-AUGMENTED GENERATION

Dieses Prinzip hat einen Namen: **RAG**, kurz für Retrieval-Augmented Generation. Retrieval heißt suchen, Generation heißt antworten. Das Wort in der Mitte, Augmented, heißt „angereichert“: Deine Frage wird mit den gefundenen Textstellen angereichert, bevor die KI antwortet.

Und genau das ist Kontext-Engineering: Nicht der perfekte Prompt entscheidet, sondern welche Infos die KI hat, bevor sie antwortet.

WAS IM HINTERGRUND PASSIERT

Programmieren musst du das nicht. Wenn du die fünf Schritte kennst, verstehst du aber, warum RAG manchmal danebenliegt und wie du das verhinderst.

1. **Zerlegen**. Deine Dokumente werden in kleine Abschnitte geschnitten, meist ein paar Absätze lang. Fachwort: *Chunks*.
2. **Einsortieren**. Jeder Abschnitt bekommt eine Art Bedeutungs-Fingerabdruck, eine lange Zahlenreihe, die beschreibt, worum es darin geht. Fachwort: *Embedding*. Abschnitte mit ähnlichem Inhalt bekommen ähnliche Fingerabdrücke, auch wenn andere Wörter drinstehen. „Kündigungsfrist“ und „wie lange vorher muss ich kündigen“ landen nah beieinander.

- 3. **Suchen.** Deine Frage bekommt ebenfalls einen Fingerabdruck. Das System sucht die Abschnitte mit dem ähnlichsten Fingerabdruck und nimmt die besten paar.
- 4. **Anreichern.** Diese Abschnitte gehen zusammen mit deiner Frage an das Sprachmodell. Sinngemäß: „Hier sind fünf Textstellen. Beantworte damit die Frage.“
- 5. **Antworten.** Erst jetzt schreibt das LLM. Weil die richtige Stelle direkt vor ihm liegt, ist die wahrscheinlichste Antwort jetzt auch die richtige.

Merk dir vor allem Schritt 3. Bei großen Datenmengen liest die KI nicht alles, sondern das, was die Suche findet. Erwischt die Suche die falsche Stelle, antwortet die KI sauber formuliert auf Basis der falschen Stelle.

Video: <https://www.youtube.com/watch?v=T-D1OfcDW1M> — IBM Technology: What is Retrieval-Augmented Generation (RAG)? Kurzes Whiteboard-Video auf Englisch, das das Prinzip in wenigen Minuten zeigt.

RAG IM VERGLEICH

	PROMPT ENGINEERING	RAG	FINE-TUNING
Was sich ändert	Deine Anweisung	Die Infos, die die KI vor der Antwort bekommt	Das Modell selbst, durch Nachtrainieren
Gut für	Ton, Format, klare Aufgaben	Eigene, private oder aktuelle Fakten	Festen Stil oder Spezialaufgaben in großer Menge
Aufwand für dich	Keiner	Dateien hochladen	Hoch, eher was für Entwickler
Neue Infos einspielen	Jedes Mal neu reinschreiben	Datei austauschen, fertig	Neu trainieren

SO NUTZT DU RAG SELBST

In ChatGPT, Claude und NotebookLM ist RAG schon eingebaut. Du lädst Dokumente hoch, das Suchen übernimmt das Tool. Für deine eigenen Unterlagen ist RAG fast immer der richtige Weg. Fine-Tuning brauchst du als normaler Nutzer praktisch nie.

Die vier Wege bauen aufeinander auf. Fang mit Weg 1 an, um das Prinzip in Aktion zu sehen. Für alles, was du öfter brauchst, nimm Weg 2 oder 3. Für Recherche mit vielen Quellen und klickbaren Belegen ist Weg 4 am stärksten.

EIN PDF DIREKT IN CHATGPT HOCHLADEN

Eine einmalige Frage zu einem einzelnen Dokument.

1. Öffne chatgpt.com und melde dich an.
2. Klick im Eingabefeld unten links auf das + und wähle das Hochladen von Dateien. Such dein PDF aus.
3. Warte, bis die Datei als kleine Kachel über dem Eingabefeld erscheint.
4. Kopier den Prompt unten, ersetze die Klammer durch deine Frage und schick ihn ab.

Teste es zuerst mit einer Frage, deren Antwort du kennst. So siehst du sofort, ob die KI die richtige Stelle findet. In Claude funktioniert das genauso, auch dort lädst du über das + im Eingabefeld hoch.

DER ANTI-RATE-PROMPT

Beantworte meine Frage ausschließlich mit dem hochgeladenen Dokument.
 Zitiere die Stelle wörtlich, auf die du dich stützt, und nenne Seite oder Abschnitt.
 Wenn die Antwort nicht im Dokument steht, schreib genau: "Steht nicht im Dokument." Rate nicht.

Meine Frage: [DEINE FRAGE]

EIN CLAUDE-PROJEKT ALS DAUERHAFTE WISSENSBASIS

Du fragst immer wieder zu denselben Unterlagen: Firmenhandbuch, Kursunterlagen, Verträge, Rezeptsammlung.

1. Öffne claude.ai/projects. Alternativ fährst du mit der Maus an den linken Rand und klickst auf **Projekte**.
2. Klick oben rechts auf + **Neues Projekt**. Gib Namen und Beschreibung ein. Claude selbst sieht diese beiden Angaben nicht, sie sind nur für dich.
3. Rechts auf der Projektseite findest du das **Projektwissen**. Klick dort auf + und lade deine Dateien hoch.
4. Klick auf **Projektanweisungen festlegen**, füg die Anweisung unten ein und speichere.
5. Starte einen neuen Chat innerhalb des Projekts. Alles im Projektwissen steht jedem Chat in diesem Projekt zur Verfügung.

Der RAG-Modus: Solange dein Projektwissen klein ist, liest Claude alles komplett. Kommt es ans Limit des Kontextfensters (so heißt das Arbeitsgedächtnis der KI), schaltet Claude in bezahlten Plänen automatisch in den RAG-Modus. Laut Anthropic passt dann bis zu zehnmal so viel rein. Du erkennst den Modus an einer Anzeige im Projekt, und im Chat siehst du, wie Claude ein Suchwerkzeug für das Projektwissen benutzt.

PROJEKTANWEISUNG (CLAUDE UND CHATGPT)

Du arbeitest mit den Dateien im Projekt.

1. Such bei jeder Frage zuerst in den Projektdateien.
2. Stütz deine Antwort nur auf das, was dort steht. Nenne immer den Dateinamen und zitiere die entscheidende Stelle wörtlich.
3. Steht etwas nicht in den Dateien, sag das klar. Dann darfst du Allgemeinwissen ergänzen, markierst es aber als "Allgemeinwissen, nicht aus deinen Dateien".
4. Widersprechen sich zwei Dateien, zeig mir beide Stellen und sag, welche neuer ist.

EIN CHATGPT-PROJEKT ANLEGEN

Du nutzt lieber ChatGPT oder willst auch Google-Drive-Dateien als Quelle verknüpfen.

1. Klick in der linken Seitenleiste auf **Neues Projekt**.
2. Gib einen Namen ein und wähle Symbol und Farbe.
3. Lade im Projekt deine Dateien hoch: PDFs, Tabellen, Textdokumente oder Bilder. Du kannst auch Text direkt einfügen.
4. Öffne über die drei Punkte (•••) die **Projekteinstellungen** und füg die Projektanweisung aus Weg 2 ein.
5. Starte deinen Chat im Projekt.

Wie viele Dateien reinpassen, hängt vom Plan ab: 5 pro Projekt kostenlos, 25 mit Plus, 40 mit Pro. Wird es eng, fass mehrere kleine Dateien zu einer zusammen.

NOTEBOOKLM FÜR RECHERCHE MIT BELEGEN

Du willst Antworten, die streng aus deinen Quellen kommen, mit klickbaren Belegen. Ideal für Studium, Recherche und lange Berichte.

1. Öffne notebooklm.google.com und melde dich mit deinem Google-Konto an.
2. Klick auf **Neues Notebook erstellen** (englisch: Create new notebook).
3. Im Fenster wählst du **Quelle hochladen** (englisch: Upload a source). Möglich sind PDFs, Websites, YouTube-Videos, Audiodateien, Google Docs und Google Slides.
4. Stell deine Frage im Chat. Hinter den Aussagen in der Antwort stehen kleine Nummern. Klick drauf, und NotebookLM zeigt dir die Stelle in deiner Quelle.

NotebookLM ist darauf gebaut, aus deinen Quellen zu antworten, mit Belegen direkt im Text. Das ist RAG in seiner strengsten Form. Ein Trick: Füg die Erklärvideos am Ende dieses Guides als YouTube-Quelle ein und stell dann Fragen dazu.

DER GEGENCHECK NACH JEDER WICHTIGEN ANTWORT

GEGENCHECK

Prüf deine letzte Antwort. Welche Aussagen stehen in meinen Dateien, und welche hast du ergänzt oder geschlussfolgert?

Mach zwei Listen:

1. Aus den Dateien (mit Dateiname und wörtlichem Zitat)
2. Ergänzt oder geschlussfolgert

KONTEXT-ENGINEERING: SO FÜTTERST DU DIE SUCHE RICHTIG

Wie gut RAG funktioniert, entscheidet sich beim Hochladen, nicht beim Fragen. Fünf Regeln:

1. **Klare Dateinamen.** „Mietvertrag_Wohnung_2024.pdf“ statt „Scan_0034.pdf“. Anthropic empfiehlt das ausdrücklich, weil gut benannte Dateien leichter gefunden werden.
2. **Ein Thema pro Projekt.** Urlaubsplanung und Rezepte im selben Projekt machen die Suche ungenauer.
3. **Echter Text statt Foto.** Ein abfotografiertes Dokument ohne Textebene ist für die Suche schwer lesbar. Test: Kannst du im PDF ein Wort markieren und kopieren? Dann ist Text drin.
4. **Alte Versionen raus.** Liegen der Vertrag von 2023 und der von 2024 nebeneinander, erwischt die Suche vielleicht den alten. Lösche, was nicht mehr gilt.
5. **Konkret fragen.** „Wie lang ist die Kündigungsfrist im Mietvertrag?“ findet die Stelle. „Was steht da so drin?“ gibt der Suche nichts in die Hand.

GRENZEN, EHRlich

RAG findet nur, was drinsteht. Fehlt die Info in deinen Dateien, hilft keine Suche. Deshalb steht „Steht nicht im Dokument“ im Prompt.

Die Suche kann danebenliegen. Bei großen Wissensbasen liest die KI nur die gefundenen Abschnitte. Aufgaben wie „Fass alle 300 Seiten zusammen“ oder „Wie oft kommt X insgesamt vor?“ sind für RAG schwer, weil dafür alles gelesen werden müsste.

Zitate prüfen. Auch mit RAG kann eine Antwort eine Stelle falsch wiedergeben. Bei allem, was zählt (Verträge, Zahlen, Fristen), klickst du die Fundstelle an und liest selbst.

Kosten. Der Einstieg ist in allen drei Tools kostenlos. In Claude kannst du kostenlos bis zu fünf Projekte anlegen, den automatischen RAG-Modus gibt es aber nur in bezahlten Plänen (Pro, Max, Team, Enterprise).

Datenschutz. Lade keine Dokumente mit sensiblen Daten anderer Personen hoch, wenn du nicht sicher bist, dass du das darfst (DSGVO). Prüfe außerdem in den Einstellungen deines Tools unter Datenkontrollen bzw. Datenschutz, ob deine Inhalte zum Training genutzt werden dürfen, und schalte das aus, wenn du das nicht willst.

WENN DAS NICHT KLAPPT

Die KI antwortet aus dem Bauch statt aus der Datei: Schreib dazu „Antworte nur mit dem hochgeladenen Dokument und zitiere die Stelle.“ In Claude-Projekten hilft es, den Dateinamen in der Frage zu nennen.

Die Datei wird nicht gelesen oder die Antworten bleiben leer: Das PDF ist wahrscheinlich ein reiner Scan. Öffne es in Google Drive mit Google Docs, dann wird der Text erkannt. Lade die neue Datei hoch.

Datei-Limit erreicht: In ChatGPT mehrere kleine Dateien zu einer zusammenfassen oder ein zweites Projekt anlegen. Alternativ nimmst du für dieses Thema NotebookLM.

NotebookLM antwortet nicht: Die Info steht nicht in den Quellen, oder die Frage ist zu allgemein. Formuliere sie konkreter.

ZUM WEITERSCHAUEN

Diese Videos erklären die Grundlagen am besten. Sie sind auf Englisch. Über das Zahnrad in YouTube kannst du automatisch übersetzte deutsche Untertitel einschalten. Das letzte Video ist auf Deutsch.

Video: <https://www.youtube.com/watch?v=LPZh9BOjkQs> — 3Blue1Brown: Large Language Models explained briefly. Rund 8 Minuten, die beste visuelle Erklärung, wie ein LLM das nächste Wort vorhersagt.

Video: <https://www.youtube.com/watch?v=zYGDpG-pTho> — IBM Technology: RAG vs Fine-Tuning vs Prompt Engineering. Martin Keen zeigt, wann welche Methode passt.

Video: https://www.youtube.com/watch?v=zjkBMFhNj_g — Andrej Karpathy: Intro to Large Language Models. Knapp eine Stunde, für alle, die tiefer einsteigen wollen.

Video: <https://www.youtube.com/watch?v=oUEfTD6-SUU> — Auf Deutsch: RAG einfach erklärt für Anfänger.

HÄUFIGE FRAGEN

IST RAG DAS GLEICHE WIE DIE WEBSUCHE IN CHATGPT ODER CLAUDE?

Fast. Bei der Websuche ist das Internet die Quelle statt deiner Dateien. Das Prinzip bleibt dasselbe: erst suchen, dann antworten.

LERNT DIE KI MEINE DOKUMENTE DAUERHAFT?

Nein. RAG verändert das Modell nicht. Die Dateien liegen im Projekt und werden bei jeder Frage neu durchsucht. Löscht du eine Datei, ist das Wissen weg.

BRAUCHE ICH RAG FÜR EIN DOKUMENT MIT ZWEI SEITEN?

Nein. Kurze Dokumente liest die KI komplett. RAG lohnt sich, wenn deine Unterlagen zu groß sind, um sie jedes Mal ganz zu lesen. In Claude-Projekten entscheidet Claude das automatisch.

WAS IST DER UNTERSCHIED ZWISCHEN PROMPT-ENGINEERING UND KONTEXT-ENGINEERING?

Prompt-Engineering ist die Formulierung deiner Anweisung. Kontext-Engineering ist alles, was die KI vor der Antwort sieht: Anweisung, Dateien, frühere Nachrichten, Suchergebnisse. RAG ist das wichtigste Werkzeug dafür.

AUF EINEN BLICK

RAG — Retrieval-Augmented Generation: erst suchen, dann antworten.

Halluzination — Eine selbstbewusste Antwort, die falsch ist.

Kontextfenster — Das Arbeitsgedächtnis der KI, also alles, was sie in einem Chat gleichzeitig sehen kann.

Chunk — Ein kleiner Abschnitt deines Dokuments.

Embedding — Der Bedeutungs-Fingerabdruck eines Abschnitts, mit dem die Suche passende Stellen findet.

Kontext-Engineering — Der KI die richtigen Infos geben, bevor sie antwortet.

HINWEIS

Hinweis: Dieser Guide dient zu Informations- und Bildungszwecken. Alle Inhalte spiegeln den Stand bei Erstellung wider. Oberflächen, Funktionen und Preise der genannten Tools können sich ändern. Prüfe Ergebnisse von KI-Tools immer selbst nach. @GPTMarlon übernimmt keine Haftung für Entscheidungen auf Grundlage dieser Inhalte. Dieser Guide stellt keine Anlage-, Steuer-, Rechts- oder Gesundheitsberatung dar. Lass Ergebnisse bei Bedarf fachlich prüfen.