Rico Eberle · LEARNING NUGGETS

LEARNING NUGGET E-LEARNING & KI 1. OKTOBER 2026

RAG erklärt: Deine KI hat das Handbuch nie gelesen

Ein KI-Assistent mit RAG sieht nur ein paar Schnipsel deiner Dokumente; passt das Material ins Kontextfenster, gibst du es oft besser ganz hinein.

Von Rico Eberle, E-Learning-Experte für Weiterbildung und Gemeinderat in Dübendorf

Video zum Beitrag

Dieses Learning Nugget gibt es als Video mit Transkript (1:47 Min.). QR-Code scannen oder Link öffnen:

ricoeberle.ch/blog/rag-was-das-llm-sieht/

Viele KI-Assistenten für eigene Dokumente haben das Handbuch nie gelesen. Sie sehen nur ein paar Schnipsel daraus, ausgewählt von einer Suche. Das Verfahren heisst RAG. Wer es versteht, kann einschätzen, wann solche Assistenten verlässlich antworten und wann nicht.

So funktioniert RAG

RAG steht für Retrieval Augmented Generation, also eine Textgenerierung, die durch eine Suche ergänzt wird. Beschrieben haben das Verfahren Patrick Lewis und Kollegen 2020.

Zuerst kommt eine Vorbereitung. Alle Dokumente werden in kleine Schnipsel zerlegt. Jeder Schnipsel bekommt einen Zahlencode für seine Bedeutung und landet in einer Datenbank. Bei jeder Frage folgen dann drei Schritte:

1. **Frage übersetzen:** Auch deine Frage wird in so einen Zahlencode übersetzt.
2. **Suchen:** Die Datenbank sucht die Schnipsel mit dem ähnlichsten Code, meist nur eine Handvoll.
3. **Antworten:** Erst jetzt kommt das Sprachmodell ins Spiel. Es bekommt deine Frage und diese paar Schnipsel und formuliert daraus die Antwort.

Das Video vereinfacht an einigen Stellen. In der Fachsprache heissen die Schnipsel Chunks, der Zahlencode Embedding und die Datenbank Vektordatenbank. Auch das Embedding entsteht mit einem neuronalen Modell, einem sogenannten Embedding-Modell; das ist aber kein Sprachmodell, das Antworten formuliert. Manche Systeme setzen zusätzlich ein Sprachmodell ein, etwa um die Frage umzuformulieren oder die gefundenen Schnipsel nachzusortieren. Wie viele Schnipsel das Modell bekommt, hängt vom System ab.

Das Modell sieht nur, was die Suche findet

Das Sprachmodell sieht nie das ganze Handbuch, nur das, was die Suche vorher ausgewählt hat. Fehlt der passende Schnipsel, fehlt dem Modell die Grundlage. Dann rät es, oder es meldet, dass es nichts findet. Zusammenhänge über mehrere Kapitel gehen leicht verloren, weil jeder Schnipsel für sich steht.

Anthropic beschreibt das Problem 2024 so: Herkömmliche RAG-Systeme entfernen beim Zerlegen den Kontext, und deshalb finden sie die relevante Information oft nicht. Ein Satz über ein Umsatzwachstum sagt allein nicht, welche Firma und welches Quartal gemeint ist. Mit zusätzlichem Kontext pro Schnipsel liessen sich solche Fehlgriffe in Anthropics Tests deutlich verringern, aber nicht ganz vermeiden.

Die Alternative: alles ins Kontextfenster

Das Kontextfenster ist alles, was ein Modell beim Antworten berücksichtigen kann, eine Art Arbeitsgedächtnis. Aktuelle Claude-Modelle fassen laut Anthropic eine Million Tokens. Tokens sind die Stücke, in die ein Sprachmodell Text zerlegt: Wörter, Wortteile oder Satzzeichen.

Wie viele Seiten das sind, hängt vom Text ab. Anthropic setzt 200’000 Tokens mit etwa 500 Seiten gleich; hochgerechnet ergibt eine Million Tokens rund 2500 Seiten. Deutsche Texte brauchen etwas mehr Tokens pro Wort, darum ist das ein Richtwert.

Gibst du alles direkt hinein, hat das Modell wirklich alles vor sich. Studien zeigen, dass das oft besser ist. Zhuowan Li und Kollegen haben 2024 beide Ansätze auf mehreren öffentlichen Datensätzen verglichen: Mit genügend Ressourcen schnitten Modelle mit langem Kontext im Durchschnitt besser ab als RAG. Auch Anthropic empfiehlt, eine Wissensbasis unter 200’000 Tokens einfach ganz in die Anfrage zu geben.

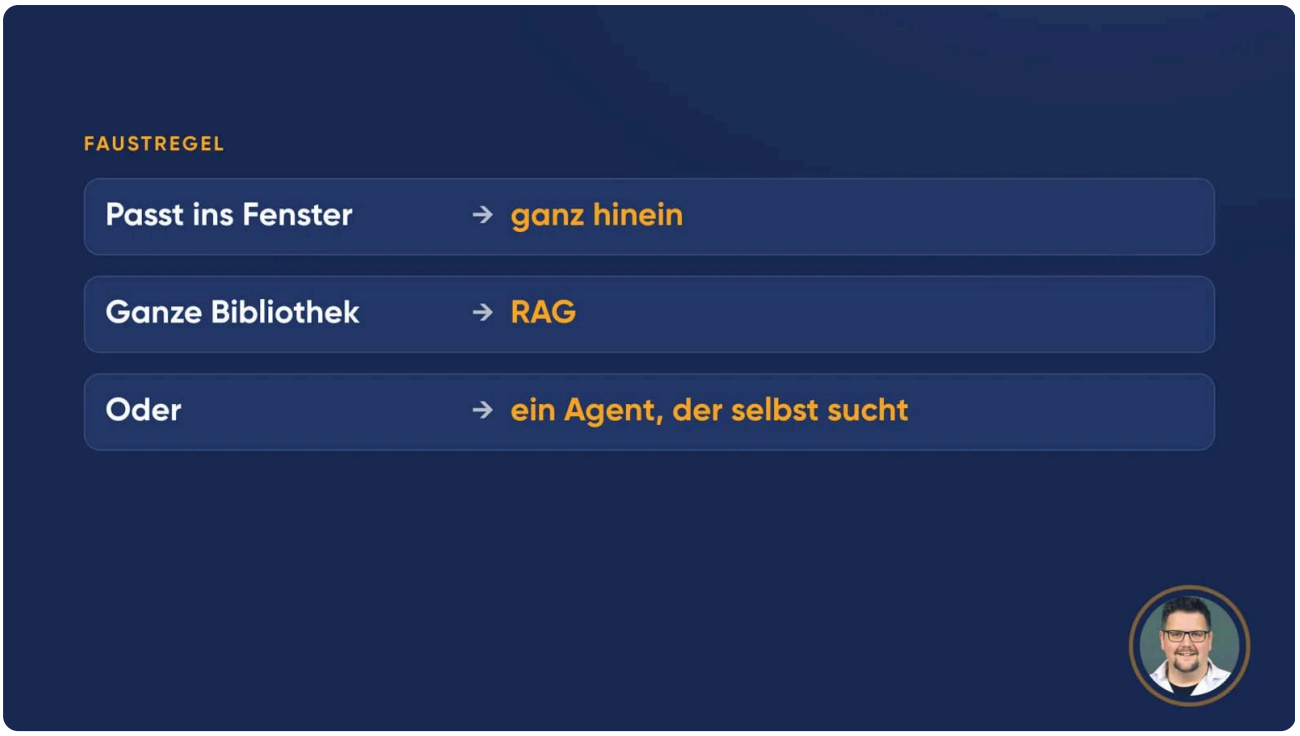
Das grosse Fenster hat Grenzen

- **Es skaliert nur bis zur Grenze des Fensters.** Eine ganze Bibliothek passt nicht hinein.
- **Jede Frage kostet mehr und dauert länger.** Abgerechnet wird pro Eingabe-Token, und das ganze Material geht bei jeder Frage mit. Zwischenspeichern (Prompt Caching) senkt die Kosten, aber nicht auf das Niveau von RAG.
- **Je voller das Fenster, desto eher übersieht das Modell Details.** Nelson Liu und Kollegen zeigten 2023, dass Modelle Informationen in der Mitte eines langen Textes schlechter nutzen als am Anfang oder am Ende («Lost in the Middle»). Chroma testete 2025 achtzehn Modelle und fand, dass die Leistung mit wachsender Eingabe nachlässt, auch bei einfachen Aufgaben. Anthropic nennt das Phänomen in der eigenen Dokumentation «context rot».

RAG hat dagegen echte Vorteile. Es ist günstig und schnell. Und weil bekannt ist, welche Schnipsel verwendet wurden, kann es zeigen, woher eine Antwort stammt. Schon Lewis und Kollegen nannten die fehlende Herkunftsangabe als Schwäche von Sprachmodellen ohne Suche.

Die Faustregel

Passt es ins Fenster, gib es ganz hinein. Ist es eine ganze Bibliothek, braucht es RAG. Oder einen Agenten, der selbst gezielt sucht.



Ein Agent wählt nicht vorab Schnipsel aus. Er sucht mit Werkzeugen selbst, prüft, was er findet, und sucht bei Bedarf weiter. Boris Cherny vom Team hinter Claude Code schrieb 2026, frühe Versionen des Werkzeugs hätten RAG mit einer lokalen Vektordatenbank genutzt; die agentische Suche funktioniere aber in der Regel besser und sei einfacher.

Antworten an der Quelle prüfen

Wer mit KI-Assistenten für eigene Unterlagen arbeitet, etwa für Kursdossiers oder Reglemente, kann daraus einiges ableiten.

Zeigt ein Assistent an, aus welchem Dokument eine Antwort stammt, lohnt sich der Blick dorthin. Das ist eine Stärke von RAG.

«Nichts gefunden» heisst nicht «gibt es nicht». Vielleicht hat die Suche den passenden Schnipsel nur nicht erwischt. Es hilft, die Frage mit anderen Begriffen zu wiederholen.

Bei Fragen über mehrere Kapitel ist Vorsicht angebracht. Vergleiche, Zusammenfassungen ganzer Dokumente oder Widersprüche zwischen Abschnitten gehen mit RAG leicht verloren. Passt das Dokument ins Fenster, ist es oft besser, es ganz mitzugeben.

Wie zuverlässig ein bestimmter Assistent ist, zeigt ein kurzer Test mit einem Dokument, das man gut kennt. Drei Fragen genügen: eine Detailfrage, deren Antwort in einem einzigen Abschnitt steht, eine Frage, die zwei Kapitel verbindet, und eine Frage, deren Antwort nicht im Dokument steht. Der Vergleich mit dem Original zeigt rasch, wo der Assistent verlässlich ist. Das Ergebnis lässt sich auch mit Teilnehmenden gemeinsam auswerten.

Was Modell, Sprachmodell, Chat und Agent genau unterscheidet, zeigt der Beitrag Modell, LLM, Chat, Agent: vier Begriffe, ein Bild.

Wer weiss, was das Modell sieht, weiss auch, was es nicht wissen kann.

Quellen

1. Lewis et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (2020)
   https://arxiv.org/abs/2005.11401
2. Anthropic: Contextual Retrieval (2024)
   https://www.anthropic.com/news/contextual-retrieval
3. Anthropic: Context windows, Claude-Dokumentation (2026)
   https://platform.claude.com/docs/en/build-with-claude/context-windows
4. Li et al.: Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach (2024)
   https://aclanthology.org/2024.emnlp-industry.66/
5. Liu et al.: Lost in the Middle: How Language Models Use Long Contexts (2023)
   https://arxiv.org/abs/2307.03172
6. Hong, Troynikov & Huber (Chroma): Context Rot (2025)
   https://www.trychroma.com/research/context-rot
7. Boris Cherny: Beitrag auf X zur Suche in Claude Code (2026)
   https://x.com/bcherny/status/2017824286489383315
8. Anthropic: Building effective agents (2024)
   https://www.anthropic.com/engineering/building-effective-agents

ÜBER MICH

Rico Eberle

Ich arbeite im E-Learning, bin Betriebsökonom FH und Gemeinderat in Dübendorf und präsidiere den Stiftungsrat der Stiftung Weiterbildungskurse Dübendorf (WBK). Die Learning Nuggets mache ich allein: Thema, Recherche, Text und Video. Fragen und Widerspruch sind willkommen.

Kontakt: ricoeberle.ch · LinkedIn: linkedin.com/in/rico-eberle

LIZENZ

Text, Transkript und Video von Rico Eberle stehen unter der Lizenz CC BY 4.0 (ausgenommen Musik und Soundeffekte des Videos). Du darfst die Inhalte teilen und bearbeiten, auch im Unterricht, wenn du die Quelle nennst.
creativecommons.org/licenses/by/4.0/deed.de

ZITIERVORSCHLAG

«RAG erklärt: Deine KI hat das Handbuch nie gelesen», Rico Eberle, ricoeberle.ch, 2026, https://ricoeberle.ch/blog/rag-was-das-llm-sieht/, CC BY 4.0