HANDOUT FORTGESCHRITTEN KI-ARCHITEKTUR

RAG verstehen

Warum KI aus deinen Dokumenten antwortet — und wann das wirklich hilft

4 Themenblöcke · April 2026 · Version 2.0

Was ist RAG — und was passiert da eigentlich?

Das Prinzip hinter KI die aus deinen Quellen antwortet

> RAG steht für Retrieval-Augmented Generation. Klingt technisch — das Prinzip ist einfach: die KI sucht zuerst in deinen Dokumenten, dann antwortet sie. Statt aus dem Trainingsgedächtnis zu raten, antwortet sie aus dem was du ihr gegeben hast.

WAS IM HINTERGRUND PASSIERT

1
Dokumente hochladen — werden in kleine Abschnitte zerlegt (Chunks)
2
Bedeutung wird gespeichert — jeder Abschnitt wird als Vektor (Bedeutungs-Koordinate) indexiert
3
Du stellst eine Frage — das System sucht die bedeutungsähnlichsten Abschnitte
4
KI antwortet aus den Fundstellen — mit Quellenangabe, nicht aus dem Trainingsgedächtnis

WAS DAS BEDEUTET

Das Modell „lernt" deine Dokumente nicht — es sucht darin. Der Unterschied ist wichtig: du kannst Dokumente jederzeit austauschen, ergänzen oder entfernen. Das Modell selbst bleibt unverändert.

> *„ RAG ist kein Feature — es ist ein Architektur-Prinzip. Die KI antwortet aus deinen Quellen statt zu raten. “*

RAG vs. PDF im Chat — der entscheidende Unterschied

Warum „einfach hochladen" nicht dasselbe ist

> PDF in ChatGPT hochladen fühlt sich wie RAG an — ist es aber nicht. Es ist ein temporärer Upload in das Kontextfenster. Das skaliert nicht, hat keine Quellenangaben und ist nach dem Chat weg.

DIREKTER VERGLEICH

PDF-UPLOAD IM CHAT

Temporär — nach dem Chat weg
Begrenzt durch Kontextfenster (~100–200 Seiten max.)
Kein Quellenverweis
Skaliert nicht (1 Dokument pro Chat)
Bei langen Dokumenten: Mitte wird vergessen

RAG-SYSTEM (Z.B. NOTEBOOKLM)

Dauerhaft — Dokumente bleiben
Skaliert auf 1.000+ Dokumente
Quellenangabe pro Antwort
Mehrere Dokumente gleichzeitig durchsuchbar
Neue Dokumente jederzeit ergänzbar

> **Wichtig:** RAG ist nicht immer besser. Für schnelle Einzelaufgaben (einen Text überarbeiten, eine E-Mail schreiben) braucht es kein RAG. RAG entfaltet seinen Wert bei wiederkehrenden Fragen an ein Dokumenten-Corpus.

> *„ PDF im Chat = Kurzzeitgedächtnis. RAG-System = Bibliothek mit Suchfunktion. “*

RAG und Halluzinationen — was es löst und was nicht

Warum RAG hilft, aber keine Garantie ist

> Halluzinationen entstehen wenn KI keine Quelle hat — und trotzdem antwortet. RAG gibt der KI einen Anker. Aber der Anker ist nur so gut wie die Dokumente die du reingibst.

WARUM HALLUZINATIONEN ENTSTEHEN

1
KI-Modelle vervollständigen Muster — sie berechnen welches Wort wahrscheinlich als nächstes kommt
2
Ohne Quelle: Mustervervollständigung — die KI „erfindet" eine plausibel klingende Antwort
3
Mit RAG: Anker statt Raten — die KI antwortet aus gefundenen Abschnitten, nicht aus Wahrscheinlichkeit

WAS RAG LÖST — UND WAS NICHT

RAG HILFT BEI

Fragen zu internen Dokumenten
Quellennachweis (du kannst prüfen)
Aktualität (Training veraltet, Docs nicht)
Transparenz: „Wo steht das?"

RAG LÖST NICHT

Schlechte Dokumente → schlechte Antworten
Fragen außerhalb des Dokuments
Widersprüchliche Dokumente
Halluzinationen komplett (nur reduziert)

> **Praxistest:** Frage das RAG-System nach etwas das nicht in den Dokumenten steht. Ein gutes System sagt: „Dazu habe ich keine Information in deinen Quellen." Ein schlechtes System erfindet eine Antwort — mit Quellenangabe auf eine Stelle die das gar nicht sagt. Quellenverweis ≠ richtige Antwort.

„ Garbage in, garbage out. RAG macht KI zuverlässiger — aber nur so zuverlässig wie die Dokumente dahinter. “

Wann RAG einsetzen — und typische Use Cases

Praktische Orientierung für den Alltag

WANN RAG SINNVOLL IST

Sinnvoll

- Interne Dokumente regelmäßig befragen
- Quellennachweis ist wichtig (Compliance, Audit)
- Aktualität zählt (Training veraltet schnell)
- Viele Dokumente, eine Wissensbasis
- Neue Kollegen onboarden mit Dokumenten-Chatbot

Bringt wenig

- Brainstorming & kreative Aufgaben
- Information nicht in Dokumenten
- Einmalige Aufgaben (schneller im Chat)
- Allgemeinwissen-Fragen
- Wenn Aktualität egal ist

TYPISCHE USE CASES BEI BOSCH

Use Case	Tool heute	Was es löst
Protokoll-Assistent	NotebookLM	„Was haben wir letzten Monat entschieden?" — ohne manuelles Suchen
Interne Wissensdatenbank	NotebookLM	Neue Kollegen fragen Handbücher und Richtlinien direkt ab
Produkthandbuch-Chatbot	Gemini + Drive Connector	Service-Mitarbeiter finden Antworten ohne Ticket zu öffnen
Richtlinien-Assistent	Custom GPT mit Knowledge	HR/Compliance-Fragen ohne Weiterleitung an Experten
Meeting-Wissensspeicher	NotebookLM	Alle Meetings eines Projekts durchsuchbar — „Was war der Stand zu X?"

DEIN NÄCHSTER SCHRITT

Schritt 1: NotebookLM öffnen → notebooklm.google.com
Schritt 2: Ein Dokument hochladen das du regelmäßig nutzt
Schritt 3: Eine Frage stellen die du normalerweise manuell recherchierst
Schritt 4: Quellenangabe prüfen — stimmt sie mit dem Dokument überein?

> *„ Der beste Einstieg in RAG ist ein Dokument das du wirklich brauchst — und eine Frage die du dir wirklich stellst. “*