HANDOUT ADVANCED KI-SPRECHSTUNDE

RAG verstehen — Advanced

Wie RAG funktioniert, warum Halluzinationen kein Fehler sind und wann sich RAG bei Bosch lohnt

6 Materialien · Erstellt: März 2026 · Version 1.0

Warum Sprachmodelle halluzinieren

Kein Bug, sondern Beiprodukt des Grundprinzips

> Ein Sprachmodell sagt das wahrscheinlichste nächste Wort vorher. Wenn es die Antwort nicht kennt, erfindet es die plausibelste. Das ist keine Fehlfunktion, das ist das Grundprinzip.

DAS GRUNDPRINZIP

Das Modell berechnet für jeden möglichen nächsten Token eine Wahrscheinlichkeit. "Die KI ist …" → *leistungsfähig* (42%), *schnell* (18%), *teuer* (11%). Es wählt basierend auf diesen Wahrscheinlichkeiten, nicht auf Faktenwissen.

DREI DINGE DIE DARAUS FOLGEN

- [ ] **Das Modell weiß nicht, was es nicht weiß**
  Es hat kein Konzept von Unsicherheit. Der Ton ist immer gleich selbstsicher, egal ob die Antwort korrekt ist oder erfunden.
- [ ] **Plausibilität ≠ Wahrheit**
  Das Modell optimiert darauf, überzeugend zu klingen. "McKinsey-Studie 2023, 47,3 % Produktivitätssteigerung" klingt belegt, kann aber komplett erfunden sein.
- [ ] **Trainings-Cutoff begrenzt das Wissen**
  Das Modell kennt nur Texte bis zu seinem Trainingsdatum. Aktuelle Entwicklungen, interne Dokumente, Änderungen seit dem Cutoff existieren für das Modell nicht.

SCHEINPRÄZISION ERKENNEN

BESONDERS ANFÄLLIG

Prozentzahlen mit Nachkommastellen, Studientitel mit Erscheinungsjahr, Rankings, historische Daten mit exakten Jahreszahlen, Unternehmenskennzahlen, Gesetzesinterpretationen. Faustregel: Je präziser eine KI-Aussage klingt, desto kritischer prüfen.

„Sicherheit im Ton ist kein Indikator für Richtigkeit des Inhalts. Das ist der Unterschied zwischen einem KI-Anfänger und einem kritischen KI-Nutzer.“

Was ist RAG und wie funktioniert es

Erst suchen, dann antworten: Das Prinzip in 4 Schritten

> RAG steht für Retrieval Augmented Generation. Statt aus dem Gedächtnis zu antworten, sucht das Modell zuerst in deinen Dokumenten. Das Ergebnis: Antworten mit Quellenangabe.

SO FUNKTIONIERT ES

1. **Chunking** — Dokumente werden in sinnvolle Abschnitte zerlegt
2. **Embedding** — Die Bedeutung jedes Abschnitts wird als Zahlenvektor gespeichert
3. **Retrieval** — Bei deiner Frage werden die bedeutungsähnlichsten Abschnitte gefunden
4. **Generation** — Das Modell bekommt Frage + Fundstellen und antwortet mit Quellenangabe

WARUM "AUTO" AUCH "FAHRZEUG" FINDET

Embeddings kodieren Bedeutung, nicht Wörter. Im Vektorraum liegen "Auto" und "Fahrzeug" nah beieinander, "Auto" und "Tischlampe" weit auseinander. Die Suche funktioniert über Bedeutungsähnlichkeit, nicht über Stichwort-Matching.

> *„RAG gibt dem Modell Anker in deinen Dokumenten. Es antwortet nicht mehr aus dem Gedächtnis, sondern aus deinen Quellen.“*

PDF im Chat vs. echtes RAG

Warum "Datei hochladen" kein RAG ist

> PDF hochladen fühlt sich an wie RAG, ist aber etwas grundlegend anderes. Wie ein Post-it am Monitor vs. ein Sachbearbeiter der das ganze Archiv kennt.

PDF im Chat

Dokument verschwindet nach der Session

Keine echten Quellenverweise auf Fundstellen

Ein Dokument, begrenzt durch Kontextfenster

Keine Aktualisierung, keine Versionierung

Kein Zugriff für andere Personen

RAG-System

Dokumente bleiben dauerhaft durchsuchbar

Quellenangabe mit Verweis auf Fundstelle

Hunderte Dokumente gleichzeitig abfragbar

Quellen jederzeit austauschbar und aktualisierbar

Skalierbar für viele Nutzer

WANN PDF IM CHAT REICHT

Einmalige Aufgabe an einem einzelnen Dokument. Schnelle Zusammenfassung ohne Quellenpflicht. Kein Wiederverwertungsbedarf.

Best Practices für RAG-Nutzung

5 Regeln für bessere Ergebnisse

1. KURATIEREN STATT MASSENHOCHLADEN

5 relevante Dokumente schlagen 50 ungefilterte. Die Qualität deiner Quellen bestimmt die Qualität der Antworten. Garbage in, garbage out gilt bei RAG besonders.

2. EIN NOTEBOOK PRO THEMA

Nicht ein riesiges "Alles-Notebook". Besser: "Marktanalyse Q1", "Onboarding BCS", "Compliance 2026". Je fokussierter die Quellen, desto präziser die Antworten.

3. QUELLEN GEZIELT EIN- UND AUSBLENDEN

In NotebookLM kannst du pro Frage bestimmte Quellen aktivieren oder deaktivieren. "Vergleiche nur Studie A und B" liefert bessere Ergebnisse als "suche in allem".

4. BESSERE PROMPTS FÜR RAG

Schwach	Stark
"Fasse das zusammen"	"Welche Risiken werden nicht adressiert?"
"Was steht da drin?"	"Wo widersprechen sich Quelle A und B?"
"Erkläre mir das"	"Welche Annahmen könnten bald nicht mehr gelten?"

5. AI READABILITY BEACHTEN

Gescannte PDFs, Tabellen als Screenshots, verschachtelte Layouts: alles schlecht für Sprachmodelle. Wenn die Zusammenfassung seltsam ausfällt, ist es oft das Format, nicht das Modell.

> *„RAG ist am stärksten wenn du es für Lücken, Widersprüche und Risiken nutzt, nicht nur für Zusammenfassungen.“*

Was RAG löst und was nicht

Grenzen kennen, Verifikation einplanen

> RAG reduziert Halluzinationen deutlich. Aber Quellenangabe heißt nicht automatisch richtige Antwort. Schlechte Dokumente ergeben schlechte Antworten, und wenn die Info fehlt, wird trotzdem geraten.

Was RAG löst

Antworten basieren auf deinen Dokumenten
Quellenangabe zeigt wo die Info steht
Aktualität steuerbar (Quellen austauschbar)
Halluzinationsrisiko deutlich reduziert
Skalierbar für Teams und Abteilungen

Was RAG nicht löst

Quellenangabe ≠ richtige Interpretation
Widersprüchliche Dokumente → widersprüchliche Antworten
Schlechte Chunks → unsichtbare Qualitätsprobleme
Fehlende Info wird weiterhin erfunden
Kritisches Denken bleibt deine Aufgabe

VERIFIKATIONS-WORKFLOW

1. Zahl, Studie oder Fakt aus der KI-Antwort kopieren
2. Originalquelle oder Google Scholar prüfen
3. Nicht auffindbar? Nicht verwenden.
4. Bei kritischen Aussagen: zweite unabhängige Quelle suchen

PROMPT-GEGENMITTEL

Nenne nur Zahlen die du aus verifizierbaren Quellen kennst. Wenn du dir nicht sicher bist, sag das explizit statt zu raten. Markiere wo dein Wissen veraltet sein könnte.

> *„Du bist nicht fertig, wenn die KI fertig ist. Auch mit Quellenangabe nicht.“*

RAG-Tools bei Bosch und Datenschutz

Was ihr heute nutzen könnt und was rein darf

VERFÜGBARE TOOLS

📓

NotebookLM

PDFs hochladen, Fragen stellen, Quellen sehen. Audio-Feature erzeugt Podcasts aus Dokumenten.

Nur unkritische Dokumente

🔗

Gemini + Drive

OneDrive/Drive-Connector. Interne Dokumente bleiben in der Bosch-Infrastruktur.

Bosch-konform

Copilot + SharePoint

RAG über eure SharePoint-Struktur. Integriert in Teams, Outlook und Word.

Bosch-konform

DATENSCHUTZ-ENTSCHEIDUNGSHILFE

- [ ] **NotebookLM = Google-Dienst**

  Keine vertraulichen internen Daten. Keine personenbezogenen Informationen. Keine Dokumente unter Geheimhaltung. Nur öffentliche oder selbst erstellte Inhalte ohne Vertraulichkeitsstufe (max SC 2).

- [ ] **Gemini Enterprise = Bosch-konform**

  Interne Dokumente bleiben in der Bosch-Infrastruktur. Die sichere Variante für alles was bei NotebookLM nicht rein darf.

- [ ] **Die Faustregel**

  Würde ich dieses Dokument einem externen Dienstleister per E-Mail schicken? Wenn nein, gehört es nicht in NotebookLM.

WANN RAG LOHNT, WANN NICHT

RAG lohnt sich

Wiederkehrende Fragen an denselben Corpus

Quellennachweis wichtig (Compliance, Audit)

Viele Dokumente, viele Fragende

Aktualität und Versionierung zählen

Overengineering

Einmalige Aufgabe an einem Dokument

Kleine Menge die jeder kennt

Kreatives Brainstorming ohne Quellen

PDF im Chat reicht völlig

> *„Das Sprachmodell weiß viel, aber nichts über euch. RAG gibt ihm eure Quellen. Euer Job bleibt: prüfen, ob die Antwort stimmt.“*