

Retrieval Augmented Generation

Antworten aus dem eigenen Aktenschrank statt aus dem Gedächtnis

Ein Sprachmodell weiß nur, was bis zu seinem Trainingsstichtag in den Daten stand, und über den eigenen Betrieb weiß es gar nichts. RAG dreht das um: Vor der Antwort sucht ein vorgeschaltetes System in freigegebenen Dokumenten und gibt die passenden Stellen mit. Das Modell formuliert dann nur noch aus diesem Material.

WARUM DAS MODELL ALLEIN NICHT REICHT

- **Fester Wissensstand.** Neue Gesetze, geänderte Preise, interne Abläufe: alles unbekannt.
- **Es füllt Lücken.** Statt „weiß ich nicht“ kommt eine erfundene, überzeugend formulierte Antwort.
- **Keine Herkunft.** Man sieht nicht, ob eine Aussage aus der Hausrichtlinie stammt oder aus irgendeinem Forenbeitrag im Training.

DER PROMPT HAT DREI TEILE

KONTEXT
Die gefundenen Auszüge aus den Dokumenten.

ANWEISUNG
Nur dieses Material verwenden, sonst nichts.

FRAGE
Die ursprüngliche Anfrage.

SO LÄUFT EINE ANFRAGE DURCH

- 1 **Die Frage.** Etwa: Welche Kündigungsfrist gilt im Dienstleistungsvertrag für Premiumkunden?
- 2 **Die Suche.** Ein Suchverfahren durchforstet die angebundene Wissensbasis und holt genau die Absätze heraus, die zur Frage passen.
- 3 **Der Kontext.** Die Fundstellen gehen zusammen mit der Frage an das Modell. Das Fachwissen kommt also im Moment der Anfrage dazu, nicht durch ein Neutraining.
- 4 **Die Antwort.** Ein gut eingestelltes System fasst zusammen, nennt die verwendeten Quellen und verweigert die Antwort, wenn der Kontext sie nicht hergibt.

Ändert sich eine Richtlinie, tauscht man das Dokument aus. Die nächste Antwort ist sofort die neue. Kein Warten auf ein Modell-Update.

WO ES SICH LOHNT

Einsatz	Was gefragt wird	Was dafür gepflegt sein muss
Interner Helpdesk	Urlaubsregelung, Spesen, IT-Abläufe	Eine gültige Fassung je Thema, keine Altversionen
Kundensupport	Produkte, Servicebedingungen	Die offiziellen Handbücher
Compliance	Datenschutz, Sicherheit, Qualität	Regelwerke inklusive ihrer Ausnahmen
Einarbeitung	Wer macht was, wie läuft welcher Prozess	Das verstreute Wissen aus Wiki und Laufwerk

WORAUS SO EIN SYSTEM BESTEHT

- **Oberfläche.** Ein Chatfenster im Intranet oder ein Feld auf der Website.
- **RAG-Dienst.** Das Herzstück: nimmt die Frage an und sucht die passenden Texte.
- **Sprachmodell.** Liest die Fundstellen und formuliert die Antwort.

EINE PROMPT-VORLAGE, DIE TRÄGT

MIT AUSSTIEG
„Antworten ausschließlich auf Basis des Kontexts. Wird die Frage darin nicht beantwortet, erfinde nichts, sondern sage klar, dass dir die Information fehlt. Liefern verschiedene Dokumente Widersprüchliches, mache den Widerspruch sichtbar, statt dich für eine Seite zu entscheiden.“

WELCHE DOKUMENTE HINEINGEHÖREN

- **Relevant.** Beantworten sie echte, häufig gestellte Fragen?
- **Aktuell.** Sind veraltete Fassungen wirklich gelöscht?
- **Widerspruchsfrei.** Zwei Dokumente mit unterschiedlicher Aussage erzeugen zwei unterschiedliche Antworten.

MERKSÄTZE

Ein RAG-System ist exakt so klug wie die Texte, die es liest.

Mehr Dokumente sind nicht besser. Dubletten erhöhen die Fehlerquote.

Der Prompt muss Wissenslücken erlauben: „Sage klar, wenn dir die Information fehlt.“