Der Weg zu einer vertrauenswürdigeren KI

Von Fine Tuning und naivem RAG zu Advanced RAG

Paul Baumgarten – 2B innovative / my-qanda

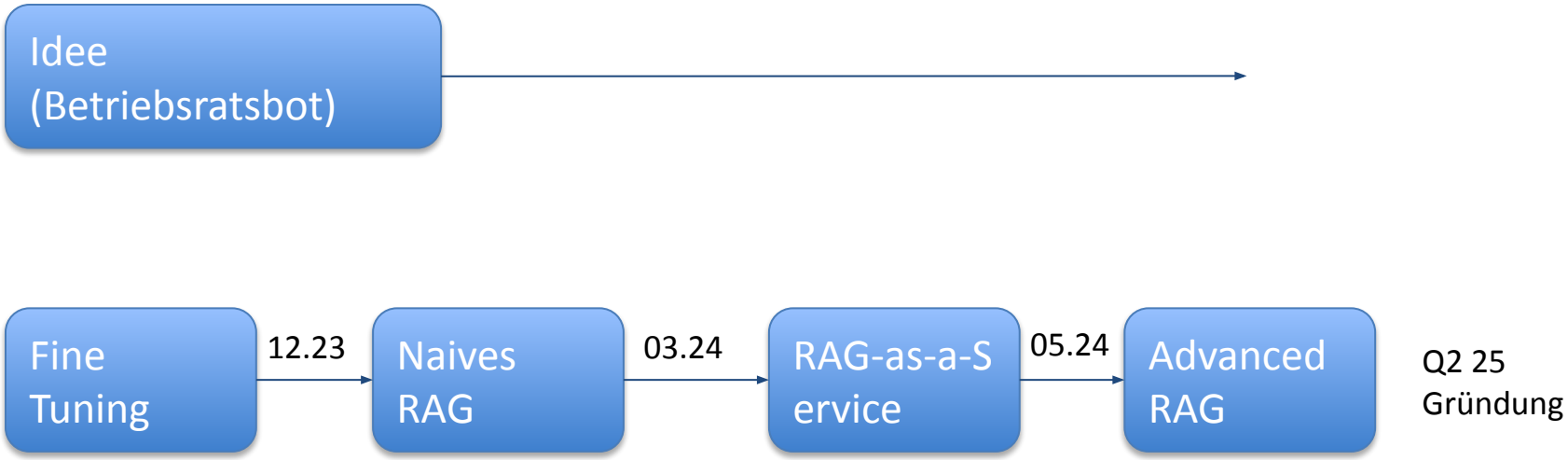
Themenübersicht

- Fine tuning - Erfahrungen (kurz)
- Retrieval Augmented Generation (länger)
  - Dokumente als Datenquelle
  - ~~Strukturierte Daten~~
- Embeddings (Optional)
- Ablauf Index/Query
- Vertrauensbildende Maßnahmen

Wie alles begann

- - Zusammenarbeit mit Hanse Betriebsratsseminare
- - Idee: Chatbot zur Unterstützung von Betriebsräten
- - Ziel: Zugänglichkeit, Effizienz, KI erlebbar machen

Von der Idee zum Produkt



Fine Tuning

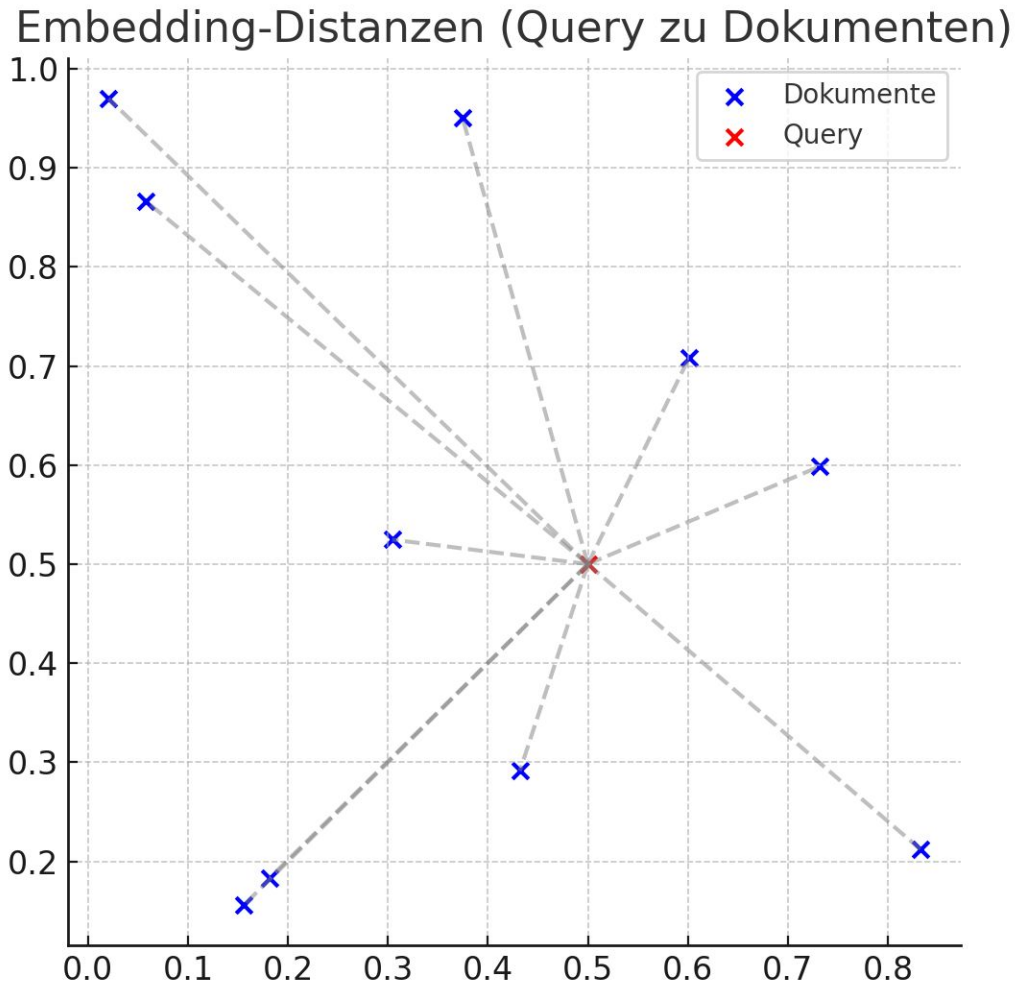
- - Aufwändig bei “flüchtigen Daten”
- - Modell muss neu trainiert werden, um Daten zu löschen
- - Halluzinationen bleiben weiterhin problematisch, unzuverlässig

Gut für den Ton, erhöht auch Genauigkeit

Retrieval-**A**ugmented **G**eneration

- - Indexing statt Training
- - Erste Ergebnisse schnell sichtbar
- - Datenzugriff transparent
- - Schnelles “Nachtrainieren”
- - Vielseitig Einsetzbar

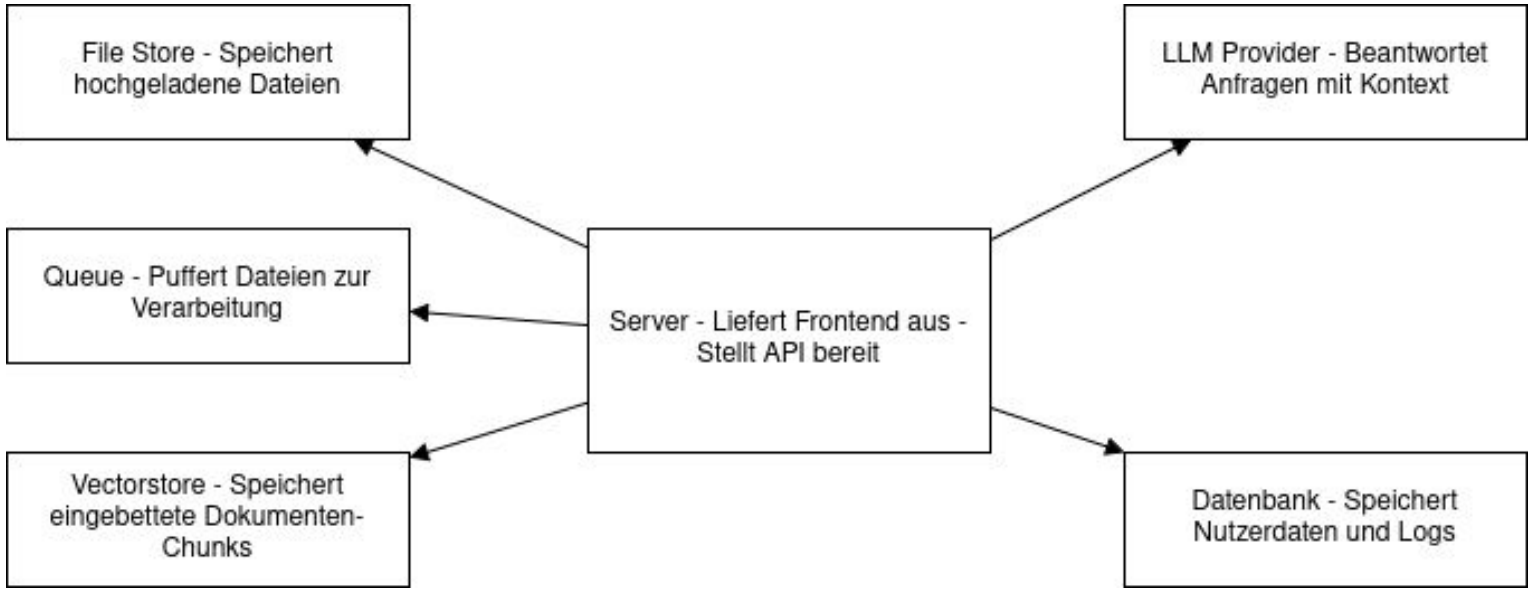
Embedding-Distanz zur Abschätzung von Halluzinationen



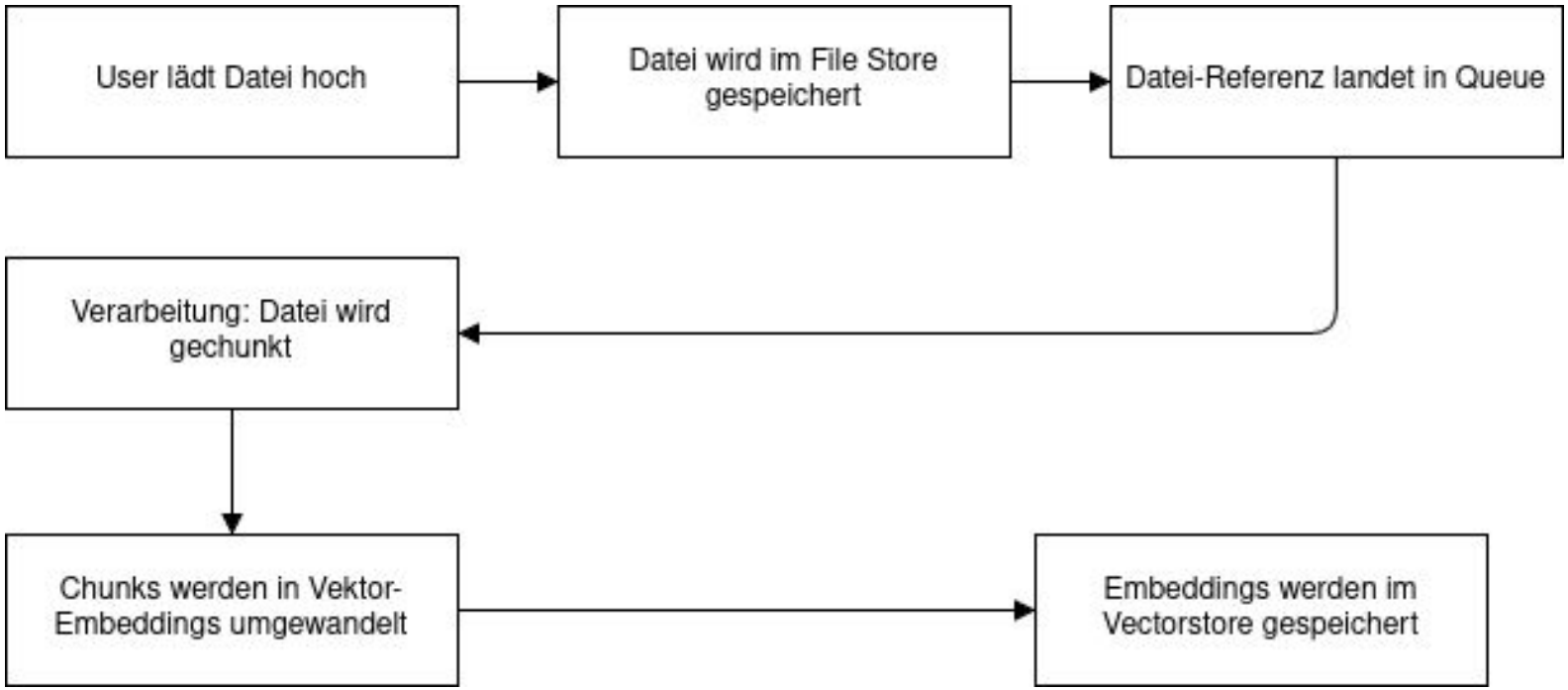
Transparente Chatbots als Ziel

- - Am Anfang „Naives RAG“: Dokumente rein, Antworten raus
- - Idee: RAG-as-a-Service mit Upload + Index API
- - Ziel: Transparente, nachvollziehbare oder keine Antwort

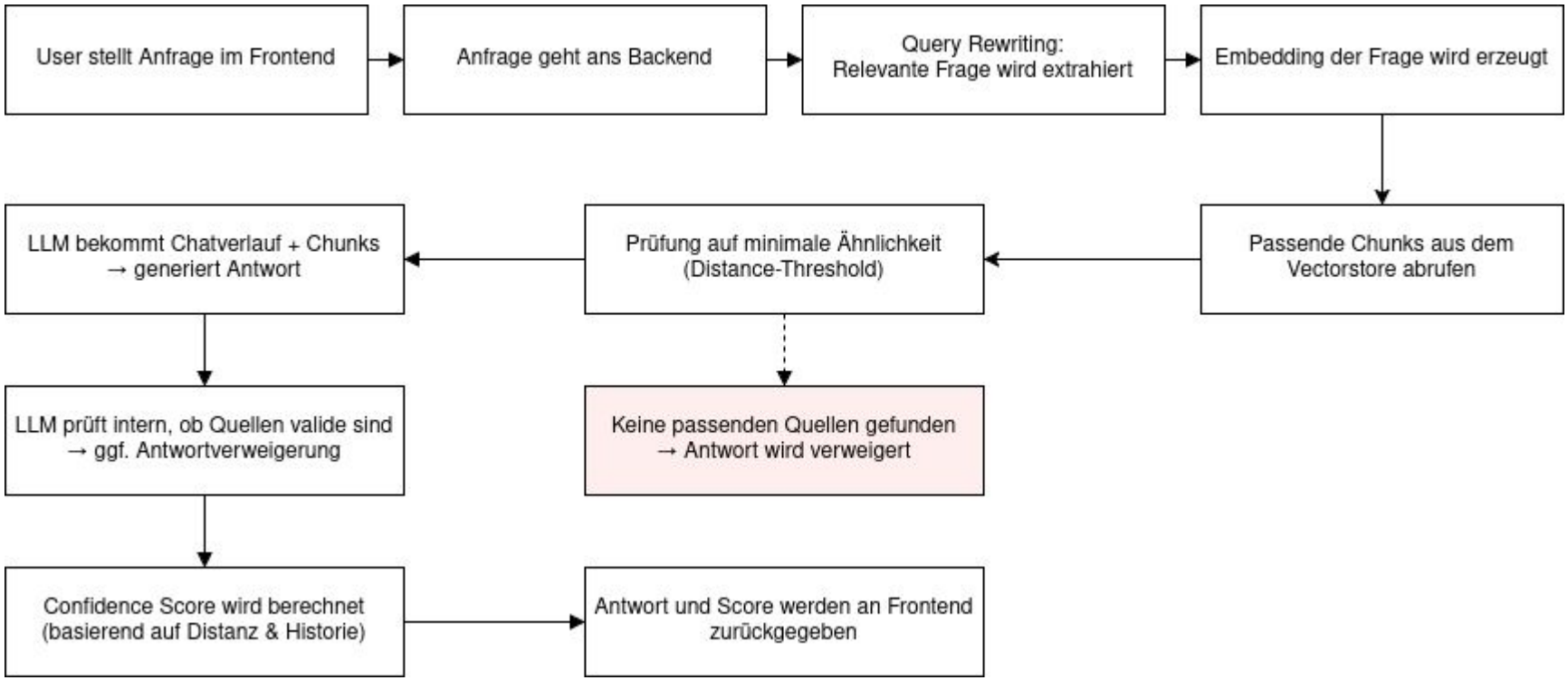
Architektur



Upload (“training”)



Anfrageverarbeitung - (Chat)



Confidence Score - Ein bisschen Transparenz

- - Embedding-Distanz als Indikator für Relevanz
- - Antwort bei zu großer Distanz verweigern
- - Transparente Bewertung für Nutzende
- - Invers zum Halluzinationsrisiko

Kein Garant für korrekte Antwort!

Fragen ?

paulbaumgarten@gmx.net - für weitere Fragen