A diagram of a process

Description automatically generated with medium confidence

**Retrieval Augmented Generation (RAG)**

* Semi-parametrisierter Informationszustand
* Verbesserung der Informationsqualität, Robustheit und Genauigkeit von LLM-Ausgaben
* Verwendung von spezifischem Wissen bei der Antwortgenerierung
* Effiziente Aktualisierung der Wissensbasis ohne Re-Training

“Naives” RAG

**Dense Retriever – Semantische Bedeutung**

Dense Passage Retriever (DPR):

* Zwei BERT-Encoder erzeugen kontextualisierte Embeddings von Frage und Passage
* Skalarprodukt der Embeddings: ähnliche Fragen und Passagen 🡪 größeres Skalarprodukt
* Encoder werden gleichzeitig trainiert Ähnlichkeit korrekt zu erfassen
* Keine lexikalische Analyse
* Benötigt große Mengen annotierter Daten

🡪 Synthetische Fragegenerierung

**Sparse Retriever – Lexikalische Analyse**

TF-IDF:

x

Termhäufigkeit skaliert nach Häufigkeit im Korpus

BM-25:

TF-IDF modifiziert durch Sättigung und Normalisierung



**Parallel Interaction**

* Interpolation von Kontext & Antwort
* Kontext-Tokens 🡪 LLM-Antwort

(Wahrscheinlichkeitsverteilung)

Beispiele:

* KNN-LM
* RETOMATON

**Internet Retriever – Suchmaschinen als Retriever**

Kommerzielle Suchmaschinen als Retriever:

* Nutzerfrage von RAG an Suchmaschine
* Liste von URLs als Suchergebnis
* Inhalt der Webseiten 🡪 Kontext zur Antwortgenerierung

Variationen:

* Verwendung von mehreren Suchmaschinen
* Bevorzugte Gewichtung von Wikipedia-Inhalte
* Gemischter Einsatz von Sparse und Dense Methoden

**Hybrid Retriever – Dense + Sparse**

* Parametrisierte Interpolation von Dense Retrieval und lexikalischen Verfahren

🡪 effektive Kombination der Vorteile

* Optimierung des Parameters mit annotierten Daten

Stufenweises Retrieval-Verfahren:

1. Bewertung aller Passagen mit BM-25
2. Dense Retrieval für Top-K Passagen

🡪 erhöhte Systemeffizienz

**Sequential Multiple Interactions**

* Für komplexere Antworten
* LLM generiert Fortsetzung als Retrieval-Query

Beispiele:

* FLARE
* RR
* REFEED

**Sequential Single Interaction**

* Einmalige Retrieval-Interaktion
* LLM erhält Top-K Passagen als Kontext

Beispiele:

* “Naives” RAG
* REALM
* Retro