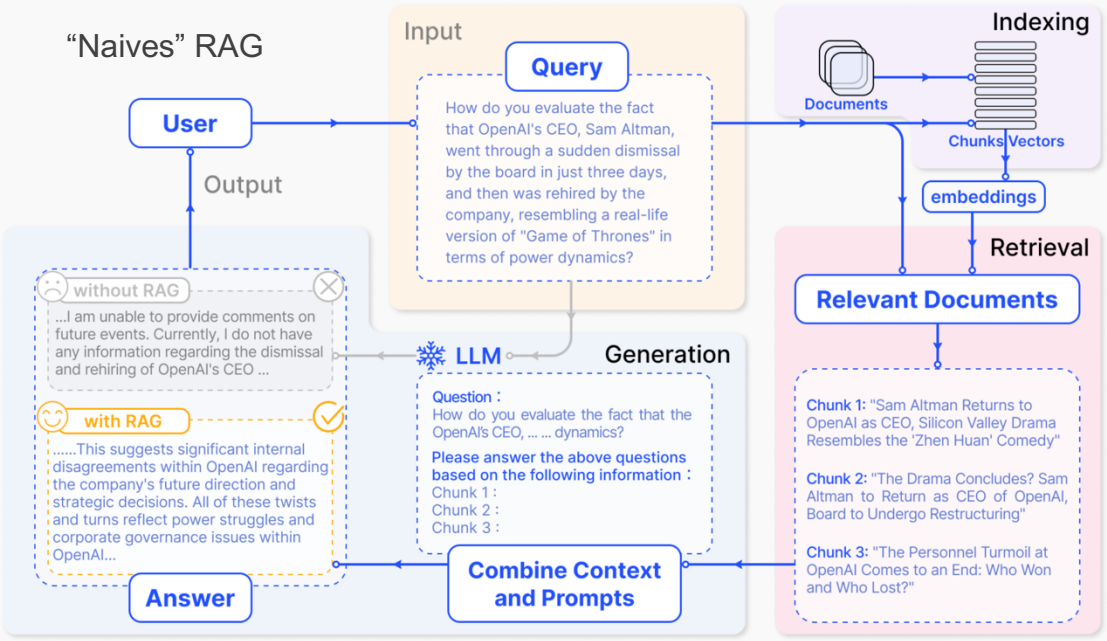
Retrieval Augmented Generation (RAG)

- 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



Sparse Retriever – Lexikalische Analyse

TF-IDF:

$$\frac{\text{Anzahl von T}}{\text{Anzahl von Wörter in P}} \text{ x } \log \frac{\text{Gesamtzahl der P im Korpus}}{\text{Anzahl der P die T beeinhalten}}$$

Termhäufigkeit skaliert nach Häufigkeit im Korpus

BM-25:

$$\sum_{i}^{n} \text{IDF}(q_i) \times \frac{\text{TF}(q_i, P) \times (k1 + 1)}{\text{TF}(q_i, P) + k1 \times \left(1 - b + b \times \frac{|P|}{\text{avg}|P|}\right)}$$

TF-IDF modifiziert durch Sättigung und Normalisierung

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

Hybrid Retriever – Dense + Sparse

- Parametrisierte Interpolation von Dense Retrieval und lexikalischen Verfahren
  → effektive Kombination der Vorteile

$$sim(q,p)^{hyb} = sim(q,p)^{DPR} + \lambda \times BM25(q,p)$$

- Optimierung des Parameters λ mit annotierten Daten

Stufenweises Retrieval-Verfahren:

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

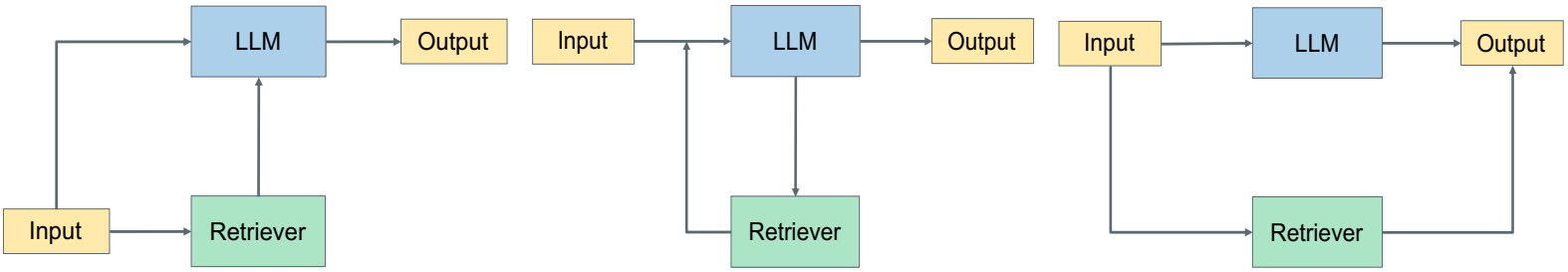
Internet Retriever – Suchmaschinen als Retriever

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



Sequential Single Interaction

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

Beispiele:

- "Naives" RAG
- REALM
- Retro

Sequential Multiple Interactions

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

Beispiele:

- FLARE
- RR
- REFEED

Parallel Interaction

- Interpolation von Kontext & Antwort
- Kontext-Tokens → LLM-Antwort (Wahrscheinlichkeitsverteilung)

Beispiele:

- KNN-LM
- RETOMATON