Was ist RAG?

Zusammenfassung der Recherche

HAW – Hamburg, Berliner Tor 21

Bachelorarbeit, Andres Toll Lück

# Einleitung

In diesem Dokument werden die Rechercheergebnisse zur Erarbeitung des Themas RAG mit entsprechenden Quellenangaben festgehalten. Für den weiteren Verlauf der Bachelorarbeit werden diese Ergebnisse in die offizielle Arbeit ggf. übertragen.

# Abstrakt

In der Bachelorarbeit sollen Kennzahlen aus vorgegebenen Dokumenten zu Bauteilnamen oder Teilenummern zugeordnet werden. Diese Bezeichnungen der Bauteile haben keine direkte Verbindung zu den Kennzahlen, wodurch Mitarbeitende einer Firma die Verbindung selbst erarbeiten müssen. Dieser Arbeitsprozess soll durch Large Language Models (LLM‘s) abgelöst werden. LLM’s sind auf einem breiten Spektrum generalisierter und unbeschrifteter Daten trainierte Modelle [2], die aufgrund von Zusammenhängen der Semantik antworten auf Eingaben (Prompts) entwickeln. Die Antwort entwickelt sich hierbei aus n-dimensionalen Vektorräumen, welche die Verbindung der Worte miteinander durch das Training darstellt. [1] Ein LLM versucht immer eine Antwort zu entwickeln und kann dementsprechend auch falsche Informationen ausgeben. Dieses Phänomen wird Halluzinieren genannt. Aus diesem Grund ist es notwendig eine Methodik zu finden, um erfundenen Kennzahlen entgegenzuwirken und nur dann eine Antwort zu erhalten, wenn eine Quelle zugrunde gelegt werden kann.

# RAG

RAG ist die Abkürzung für Retrieval-Augmented-Generation (Abruf Ergänzende Generierung). Zitat AWS Amazon: *RAG ist ein Ansatz zur Lösung einiger dieser Herausforderungen. Es leitet das LLM weiter, um relevante Informationen aus maßgeblichen, vorab festgelegten Wissensquellen abzurufen. Unternehmen haben eine bessere Kontrolle über die generierte Textausgabe, und die Benutzer erhalten Einblicke in die Art und Weise wie das LLM die Antwort generiert.* [2]

## Funktionsweise

RAG ist eine vorgeschaltete Ebene, in der relevante Informationen passend zum Prompt herausgesucht werden und anschließend mit dem Prompt zusätzlich an das LLM weitergeleitet werden. [2] So kann ein statisches LLM durch aktuelle Informationen gespeist werden und gezieltere Antworten auf Anfragen geben, welche mit Quellenangabe bewiesen werden können.

## Daten

Die Daten, aus denen diese Zusätzlichen Informationen entstammen, können sowohl aus lokalen Dokumenten entspringen oder auch aus externen Datenbanken und Informationsquellen entspringen. Die Informationen werden durch eine andere KI-Technik in numerische Darstellungen konvertiert und in einer Vektordatenbank gespeichert. Diese Daten können dann von einer generativen KI interpretiert werden. [2]

## LLM Prompterweiterung

Das RAG-Modell erweitert den Prompt für das LLM durch die vorausgegangene Anfrage und den dadurch abgerufenen Informationen aus dem vortrainierten Vektorraum auf Basis zusätzlicher Informationen. Die durch den Kontext entstandenen Zusatzinformationen werden anschließend durch Prompt-Engineering-Techniken der vorausgegangenen Anfrage hinzugefügt. Diese Erweiterung ermöglicht es dem großen LLM eine Antwort auf die Anfrage mit dem Kontext durch RAG zu formulieren und entsprechende spezielle Informationen wiederzugeben.

## Prozessablauf

In der folgenden Abbildung wird dargestellt, wie RAG über die zusätzlichen Informationen den Prompt ergänzt und dem LLM entsprechend zusätzliche Informationen liefert, damit es eine geeignete Antwort auf die Anfrage formulieren kann.

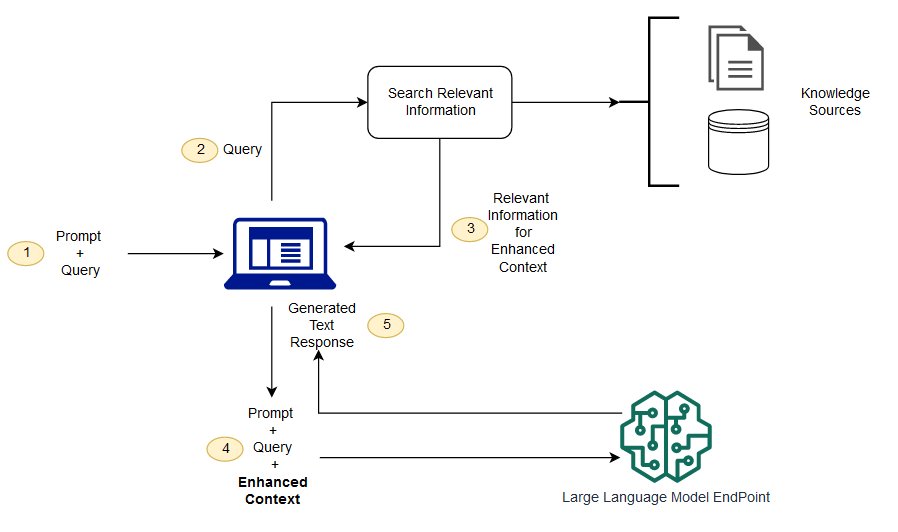


Abbildung 1 - RAG Prozess [2]

Ein klassischer Weg nach der Abbildung ist 1 zu 4 (Ohne Enhanced Context) zu 5.

# Quellenverzeichnis

Das Quellenverzeichnis ist nach folgenden Vorgaben beschriftet worden:

Buch:  
<https://www.scribbr.de/richtig-zitieren/quellenangabe-buch/>  
(28.04.2025)

Online:  
\*?

**[1] Ananthaswamy, A. (2024). Why Machines Learn: The Elegant Math Behind Modern AI, Dutton**

**[2] Was ist Retrieval-Augmented Generation (RAG)?**  
<https://aws.amazon.com/what-is/retrieval-augmented-generation/>  
(28.04.2025, 13:47)