
Empirische Analyse von RAG Evaluation Tools für betriebliche Abläufe

Bachelorarbeit zur Erlangung des akademischen Grades
Bachelor of Science
im Studiengang Allgemeine Informatik
an der Fakultät für Informatik und Ingenieurwissenschaften
der Technischen Hochschule Köln

Vorgelegt von: Leon Alexander Bartz
Matrikel-Nr.: 1114236017
Adresse: Richard-Wagner-Straße 47
50679 Köln
leon_alexander.bartz@smail.th-koeln.de

Eingereicht bei: Prof. Dr. Boris Naujoks
Zweitgutachterin: Prof. Dr. Dietlind Zühlke

Köln, 30.06.2025

Kurzfassung/Abstract

Heutzutage gewinnen LLMs wie ChatGPT und Gemini immer weiter an Beliebtheit. In Kombination mit relevanten, häufig firmeninternen Dokumenten, können Sie noch hilfreicher sein. Wenn man mithilfe von LLMs jetzt relevante Daten in einer Datenbank sucht um damit Kontextabhängige Fragen zu beantworten hat man ein RAG. Wie gut RAGs jedoch funktioniert hängt von vielen unterschiedlichen Faktoren, diese System manuell zu bewerten erfordert einen Zeit und Kostenintensiven Prozess.

Retrieval Augmented Generation Assessment (RAGAS) hat ein System entwickelt um RAGs automatisch zu bewerten. Wie gut diese automatisierte Bewertung von RAGs mithilfe von RAGAS funktioniert ist der Hauptfokus dieser Arbeit. Dabei werden besonders drei Bereiche untersucht, die generierten Fragebögen, die Bewertung und die Zuverlässigkeit bei mehreren Wiederholungen.

Es war klar zu beobachten, dass das genutzte LLM eine große Rolle spielte, insbesondere bei der Fragebogen generierung als auch bei der Bewertung. Bessere LLMs haben sowohl weniger Fehler während der generierung der Fragebögen gemacht als auch bessere Fragen generiert. Es hat sich außerdem gezeigt, dass Fehler aus den Fragebögen sich durch die Bewertung ziehen und dadurch die Bewertung negativ beeinflussen. Die Metriken und Bewertungen waren konstant über mehrere Bewertungen hinweg und es waren nur minimale Schwankungen fest zu stellen.

Zusammenfassen lässt sich sagen, dass Ragas nicht als alleiniger Faktor eingesetzt werden kann und mindestens bei der Testset generierung eine menschliche Überprüfung stattfinden sollte. Außerdem wäre ein LLM welches gegen Entgelt schnell Ergebnisse liefert zu empfehlen. Zukünftig lässt sich Ragas mit weiterentwickelten LLMs und einer verbesserten Fragebogen generierung vielleicht komplett automatisieren !TODO!

Schlagwörter/Schlüsselwörter: gegebenenfalls Angabe von 3 bis 10 Schlagwörtern. LLM, KI, RAG, Ragas, Automatisierung

Inhaltsverzeichnis

Tabellenverzeichnis	V
Abbildungsverzeichnis	VI
1 Einleitung	1
1.1 Wie funktioniert ein LLM	2
1.2 Wie funktioniert ein RAG	3
1.2.1 Vorteile von RAGs	3
1.2.2 Kompetenz des Betreibers	4
1.2.3 Datenbasis	4
1.2.4 Budget	5
1.3 Objektive Beurteilung von RAGs	5
1.4 Darstellung des Themas und der Forschungsfragen	5
1.5 Praxistauglichkeit und Herausforderungen	6
1.6 Softwaretechnische Fragestellungen	6
1.7 Rechtliche Fragestellungen	7
2 Methoden und Materialien	8
2.1 Werkzeuge	8
2.1.1 Ollama [16]	8
2.1.2 Embeddings	8
2.1.3 Vektordatenbank ChromaDB [3]	8
2.1.4 Retrieval Augmented Generation Assessment (RAGAS) [7]	9
2.1.5 Langchain [13]	9
2.2 Daten	9
2.3 Fragebögen	10
2.3.1 Fragetypen	10
2.3.2 Wissensgraph	11
2.3.3 Evaluation	12
3 Metriken	13
3.1 Retrieval Augmented Generation	13
3.1.1 Context Precision	13
3.1.2 Context Recall	14

3.1.3	Response Relevancy	14
3.1.4	Faithfulness	14
3.1.5	Context Entities Recall	15
3.1.6	Noise Sensitivity	15
3.1.7	Multimodal Faithfulness/Multimodal Relevance	16
3.2	Nvidia Metrics	16
3.2.1	Answer Accuracy	16
3.3	Natural Language Comparison	16
3.3.1	Factual Correctness	16
3.3.2	Semantic Similarity	17
3.4	Non LLM String Similarity	17
3.4.1	BLEU Score	17
3.4.2	ROUGE Score	17
3.4.3	String Presence	18
3.4.4	Exact Match	18
3.5	General Purpose	18
3.6	Andere Metriken	18
3.6.1	Summarization	18
4	Ähnliche Arbeiten	19
4.1	RAG Evaluation: Assessing the Usefulness of Ragas	19
4.2	Benchmarking Large Language Models in Retrieval-Augmented Generation	19
4.3	RAG-Bewertungsprozess	21
5	Versuche	23
5.1	Versuchsplan	23
5.1.1	Forschungsfragen	23
5.1.2	Variablen in den Versuchen	23
5.1.3	Kosten- und Zeitanalyse	25
5.1.4	Versuchsprotokoll	25
5.1.5	Bewertungskriterien für die Geschäftstauglichkeit	26
5.2	Konkretisierung der Versuche	27
5.2.1	Dokumentenverarbeitung	27
5.2.2	Testset-Generierung	27
5.2.3	Bewertung	27
6	Ergebnisse und Diskussionen	29
6.1	Ergebnisse aus den Versuchen	29
6.1.1	Generierte Fragebögen	29
6.1.2	Manuelle Auswertung der Fragebögen	30
6.1.3	Auswertung der Reports	33

6.1.4	Unterschiede über mehrere Durchläufe	37
6.1.5	Zuverlässigkeit von Metriken	39
6.1.6	Kostenberechnung	42
6.2	Abhängigkeit der Metriken untereinander	42
7	Zusammenfassungen	44
7.1	Benutzung von RAGAS	44
7.2	Fragebögen	44
7.3	Bewertung	45
7.4	Zuverlässigkeit	45
7.5	Fazit	45
7.6	Zukunftsausblick	46
7.7	Reflexion der Arbeit	46
	Literatur	47
	Anhang	50

Tabellenverzeichnis

5.1	Kombinationen aus Dokumentenanzahl und Embedding-Modell für die Versuche (X = Kombination wird getestet)	27
5.2	Kombinationen aus Dokumentenanzahl und Testset-Größe	27
5.3	Übersicht aller 24 zu generierenden Bewertungsberichte mit Abkürzungen und Wiederholungen	28
6.1	Übersicht der generierten Fragen und Fehlerquoten pro Testset für DeepSeek	30
6.2	Übersicht der generierten Fragen und Fehlerquoten pro Testset für OpenAI	30
6.3	Anzahl fehlerhafter Fragen pro Testset und Gesamtübersicht für Ollama .	32
6.4	Anzahl fehlerhafter Fragen pro Testset und Gesamtübersicht für OpenAI .	33
6.5	Verteilung der Bewertungen für DeepSeek (mit Prozentangaben)	34
6.6	Verteilung der Bewertungen für OpenAI mit Prozentangaben	34
6.7	Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für DeepSeek	37
6.8	Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für GPT-4	39
6.9	Dauer der Evaluation pro Dokumentenanzahl mit DeepSeek	41
6.10	Dauer der Evaluation pro Dokumentenanzahl mit OpenAI	41

Abbildungsverzeichnis

1.1	Struktur eines RAGs, Quelle: [14]	3
2.1	Beispiel für ein Embedding von "Mann und Frau" in einem niedrigerdimensionalen Raum.	9
4.1	Flussdiagramm des RAG-Bewertungsprozesses, das die Interaktion zwischen verschiedenen Komponenten und Modellen zeigt. Spezifische Modellnamen (z.B. gpt-4-turbo, text-embedding-3-large) sind im Haupttext beschrieben.	22
6.1	DeepSeek Ergebnis für 300 Fragen (mit Code-Dokumenten)	35
6.2	ChatGPT Ergebnis für 300 Fragen (mit Code-Dokumenten)	36
6.3	Abweichungen des Faithfulness Scores bei Code-Dokumenten.	36
6.4	DeepSeek Ergebnis für 100 Fragen (ohne Code-Dokumente)	37
6.5	ChatGPT Ergebnis für 100 Fragen (ohne Code-Dokumente)	37
6.6	Bewertung der vier Durchläufe mit DeepSeek	38
6.7	Bewertung der vier Durchläufe mit GPT-4	38
6.8	Abweichungen des Answer Relevancy Scores.	40
6.9	Abweichungen des Faithfulness Scores.	40
6.10	Abhängigkeit der Metriken voneinander (OpenAI, 300 Fragen, 400 Dokumente)	43

1 Einleitung

Im Jahr 2022 veränderte OpenAI mit ihrem browserbasierten ChatGPT (Generative Pre-trained Transformer) [17] die Welt komplett. In nur fünf Tagen erreichte ChatGPT eine Million Nutzerinnen [30] und ist aus dem Alltag vieler Menschen nicht mehr wegzudenken. Die GPT-KI (Künstliche Intelligenz) von OpenAI gehört zur Familie der Large Language Models (LLMs) oder auch Multimodal Large Language Models (MLLMs). MLLMs können neben Text weitere Datenmodalitäten wie zum Beispiel Bilder, Audio und Video verarbeiten. LLMs von anderen Anbietern wie Googles Gemini [12] und DeepSeek [4] haben mit der Qualität und den Fähigkeiten von OpenAIs GPT gleichgezogen. Mittlerweile gibt es viele Arten, LLMs zu bewerten, und ein reger Wettbewerb ist um die vielen Bewertungen entstanden.

Die GPT-Modelle von OpenAI und anderen Anbietern wie Googles Gemini [11] sind meistens nur über eine API (Programmierschnittstelle) gegen Entgelt verfügbar. Open-Source-Modelle wie DeepSeek [4] oder Metas LLAMA [31] erfreuen sich immer größerer Beliebtheit, da sie gratis auf der eigenen Hardware ausgeführt werden können.

Im Oktober 2023 kam der Verfasser das erste Mal mit Retrieval-Augmented Generation (RAGs) in Kontakt; damals war die Idee, mit Hilfe eines LLMs Fragen über mehrere firmeninterne Informationsquellen zu beantworten. Bei einem Hackathon gelang es dem Team des Verfassers, einen Prototypen (im Folgenden System genannt) zu entwickeln, der mit einem gewissen Erfolg Fragen zu firmeninternen Themen beantworten konnte.

Einer der Schritte während der Entwicklung war das ständige Testen der neuesten Änderungen. Dadurch konnte die Funktionsfähigkeit überwacht und eventuell schlechte Ergebnisse dokumentiert werden. Diese zeitintensive Aufgabe kostete uns wertvolle Zeit, welche das Team lieber in die Entwicklung investiert hätte. Gerne hätten wir unterschiedliche Prompts (Vorlagen von Fragen an ein LLM) innerhalb unseres Systems getestet oder eine automatische Überprüfung unserer neuesten Änderungen genutzt.

RAGAS [5] wurde entwickelt, um diese Probleme zu lösen. Es hat zudem das Alleinstellungsmerkmal, dass man weder eigene Fragen noch die generierten Fragen selbst beantworten muss. Sowohl die Generierung eines Fragenkatalogs (Testsets) als auch die Beantwortung der Fragen, um eine Musterlösung zu erstellen, nimmt RAGAS mithilfe von LLMs vor. Mithilfe dieses Testsets und von RAGAS eigens entwickelter Metriken, welche die wichtigsten Funktionen eines RAGs abdecken, kann eine Bewertung des Systems vorgenommen werden.

Damit benutzt RAGAS die neue LLM-Technologie, um das durch LLMs entstandene System selbst zu testen. Dies spart menschliche Ressourcen, welche zeit- und kostenintensiv sind. Wie gut LLMs für diese Aufgabe geeignet sind, ist jedoch auch aufgrund ihrer Halluzinationen fraglich.

1.1 Wie funktioniert ein LLM

Die meisten LLMs heutzutage sind sogenannte Transformer, daher kommt auch das T in GPT. Transformer sind Neuronale-Netzwerk-Modelle, welche auf dem Aufmerksamkeits-Mechanismus basieren. Das Konzept der Aufmerksamkeit wurde im Paper „Attention Is All You Need“ [32] vorgestellt; es ist einer der fundamentalen Bausteine für die heutigen LLMs. LLMs setzen sich aus vier Schritten zusammen:

- Tokenisierung (Tokenizer)
- Einbettung (Embedding)
- Berechnung der Wahrscheinlichkeit des nächsten Tokens (Vorhersage)
- Strategien zur Auswahl der Ausgabe (manchmal auch Dekodierung genannt).

TODO cite <https://www.iese.fraunhofer.de/blog/wie-funktionieren-llms/> LLMs arbeiten mit sogenannten Tokens; je nach Verfahren bestehen Token aus einzelnen Zeichen bis hin zu ganzen Wörtern. Die Aufgabe des Verwandeln der Eingabe in Tokens übernimmt der Tokenizer. <https://tiktokenizer.vercel.app/>

Damit ein Neuronales Netzwerk mit Tokens rechnen kann, müssen diese in Vektoren umgewandelt werden. Ein Umwandeln in einfache Zahlen reicht hier nicht, da wir die Fähigkeit benötigen, semantische Ähnlichkeiten zu modellieren. Das Umwandeln ist mithilfe von Embeddings möglich. Embeddings sind neuronale Netze, welche mit großen Mengen an Texten trainiert wurden, um Wörtern eine Position im höherdimensionalen Raum zu geben, welche ihre semantische Bedeutung beibehält.

Die Vektoren können jetzt in ein Neuronales Netzwerk, das meistens die Transformer-Architektur verwendet, gefüttert werden. Die Ausgabe dieses Schrittes besteht aus den Wahrscheinlichkeiten für alle möglichen nächsten Token.

Im letzten Schritt muss ausgewählt werden, welcher Token genutzt werden soll. Würde hier immer nur der wahrscheinlichste Token gewählt werden, würde man nur Texte generieren, welche das LLM während des Trainings bekommen hat. Um einen neuen Text zu generieren, werden also zufällig weniger wahrscheinliche Tokens ausgewählt. Dieser Prozess führt zu dem, was als Halluzinationen bekannt ist. Es ist ein fester Bestandteil der LLM-Architektur.

1.2 Wie funktioniert ein RAG

Bei Retrieval Augmented Generation (RAG) erweitert man den Prompt für das LLM um Suchergebnisse aus einer Dokumentensammlung, einer Datenbank, einem Wissensgraphen (Knowledge Graph) oder einer anderen Suche (z.B. Internetsuche). Das Wissen für die Antwort kommt also aus angebundenen Quellen und nicht aus dem LLM.

[14]

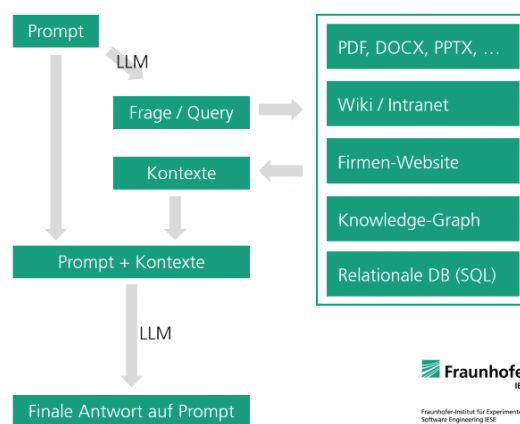


Abbildung 1.1: Struktur eines RAGs, Quelle: [14]

1.2.1 Vorteile von RAGs

Bei der Wissensabfrage durch LLMs zeigen sich unter anderem folgende Schwachstellen:

1. Im Trainingsset für die LLMs selten vorkommendes Wissen können selbst LLMs schlecht lernen. [9] [19]
2. LLMs kennen nur die verwendeten Trainingsdaten und müssen weiter trainiert werden, um neue Informationen zu erlernen.
3. Firmeninterne Dokumente sind nicht im Trainingsset, und daher können LLMs keine Fragen zu firmeninternen Daten beantworten.

Die Nutzung eines RAGs ist eine der drei Möglichkeiten, um ein LLM zu verbessern. Die anderen beiden Möglichkeiten sind Finetuning und die Verwendung eines LLMs mit einem großen Kontextfenster. Die Größe des Kontextfensters beschreibt die Menge an Text, gemessen in Tokens, die das LLM gleichzeitig berücksichtigen kann. Ein Token ist ein Teil eines Wortes, ein Wort oder ein Satzzeichen. RAGs haben neben dem Finetuning und der Nutzung eines LLMs mit großem Kontextfenster entscheidende Vorteile.

Es gibt einige Faktoren, welche die Entscheidung beeinflussen können, ob ein RAG besser für den betrieblichen Ablauf geeignet ist. Dazu gehören z. B. die Kompetenz der Betreiber des RAGs, die Art der Daten und die finanziellen Möglichkeiten des Unternehmens.

1.2.2 Kompetenz des Betreibers

Für das Finetuning von LLMs ist technisches Wissen notwendig, um die Themen Natural Language Processing (NLP), Deep Learning, Modellkonfiguration, Datenaufbereitung und Evaluierung anzuwenden. Der gesamte Prozess des Finetunings ist technisch anspruchsvoll, erfordert das Sichten der neuen Trainingsdaten und ist zudem durch die benötigte Hardware teuer.

Das Benutzen eines LLMs benötigt die geringste Kompetenz des Betreibers, da hier das LLM unverändert bleibt. Hier werden einfach die Daten inklusive der Frage an das LLM gesendet.

Während das LLM in einem RAG auch unverändert bleibt, wird es in ein System mit mehreren Komponenten eingebunden. Hier ist ein allgemeines Verständnis von LLMs und effektiven Methoden für den suchenden (Retrieval) Teil des RAGs notwendig. Zudem müssen hier manuell für jedes Dateiformat (E-Mail, PDF etc.) Anbindungen erstellt werden. Sollte ein seltener oder proprietärer Datentyp verwendet werden, muss hier eventuell eigens eine Anbindung entwickelt werden.

1.2.3 Datenbasis

Sollten die Daten dynamisch sein, wie E-Mails, ist das RAG die vorzuziehende Lösung. Dies liegt an den Eigenschaften der schnellen und kontinuierlichen Aktualisierung der Daten. Wie oben erläutert, kann es jedoch sein, dass es schlechte oder keine Unterstützung von selten verwendeten Dateiformaten gibt.

Der Prozess des Finetunings erstellt hingegen eine Momentaufnahme, die ein erneutes Training erfordert. Beim Finetuning ist es möglich, dass das Modell Muster erkennt und firmeneigene Begriffe verstehen kann. Dies ist ein deutlicher Vorteil gegenüber den anderen Methoden.

1.2.4 Budget

Das Finetuning erfordert über einen langen Zeitraum teure Rechenzeit auf Hochleistungs-GPUs. Dabei handelt es sich um spezialisierte Grafikprozessoren, deren Architektur auf die massive parallele Verarbeitung von Daten optimiert ist, was sie für die rechenintensiven Operationen des Machine Learnings, insbesondere neuronale Netze, unerlässlich macht. Zudem ist die Qualität der Daten entscheidend; ohne das vorherige Filtern der Daten durch Menschen ist dieser Prozess aktuell nicht möglich. Sollten sich die Daten als unzureichend erweisen, ist die gesamte Rechenzeit verschwendet gewesen.

Das RAG verursacht dagegen zusätzliche Kosten durch das Speichern der Daten in einer Vektordatenbank.

Die wohl kostenintensivste Methode ist die Nutzung eines LLMs mit einem großen Kontextfenster.

1.3 Objektive Beurteilung von RAGs

Je mehr Daten einem RAG zur Verfügung stehen, desto aufwendiger ist es, die Qualität des RAGs zu beurteilen. Eine Beurteilung durch Menschen müsste bei Anpassungen am RAG oder Änderungen an den Daten neu durchgeführt werden.

Tools wie RAGAS, die bereits eine automatisierte Bewertung versuchen, nutzen bei diesem Prozess unter anderem LLMs. Diese Tools generieren aus den ihnen gegebenen Daten Fragebögen, die zu einer Frage eine beispielhafte Antwort und die genutzten Stellen aus den vorher gegebenen Dokumenten beinhalten. Sollten nach diesem automatisierten Test die gewünschten Ergebnisse nicht erreicht werden, kann beispielsweise die Veröffentlichung blockiert werden.

Sowohl menschliche Bewertungen als auch die reine subjektive Bewertung durch LLMs sind jedoch nicht objektiv. Anhand mehrerer Techniken kann versucht werden, die Bewertung mithilfe von LLMs zu objektivieren. Die Metriken, welche RAGAS nutzt, versuchen z.B. die Anzahl der genannten Fakten aus den Antworten zu extrahieren, um so mit Zahlen arbeiten zu können.

1.4 Darstellung des Themas und der Forschungsfragen

In dieser Arbeit soll untersucht werden, wie gut mithilfe von aktuellen RAG-Evaluations-Tools wie RAGAS RAGs bewertet werden können. Der Fokus liegt dabei darauf, die Qualität der generierten Fragebögen und der Metriken zu bewerten.

- Können RAG-Evaluierungstools wie RAGAS aussagekräftige und kontextrelevante Fragebögen zur Bewertung von RAG-Systemen generieren?
- Ermöglichen die durch RAGAS bereitgestellten Metriken und Bewertungen eine valide und zuverlässige Einschätzung der Leistungsfähigkeit von RAG-Systemen?
- Zeigen die mit RAGAS durchgeführten Bewertungen signifikante Schwankungen, und welche Implikationen ergeben sich daraus für die Verlässlichkeit der Evaluation?

1.5 Praxistauglichkeit und Herausforderungen

Vor Beginn dieser Arbeit antizipierte der Autor bereits Schwierigkeiten bei der Bewertung von Retrieval-Augmented Generation (RAG)-Systemen mittels RAGAS:

- Die Kosten, die bei der Bewertung entstehen.
- Die Dauer für die Durchführung der Bewertung. Die Bewertung kann schneller durchgeführt werden, wenn mehr Ressourcen zur Verfügung stehen.
- Das Aufsetzen des zu testenden Systems. Dies beinhaltet eine eventuelle doppelte Speicherung der Daten und die für das Testen benötigten Aufrufe des LLMs.
- Das System muss auf dem neuesten Stand gehalten werden, da sich dieses aktuelle Thema schnell entwickelt.

1.6 Softwaretechnische Fragestellungen

In dem Artikel *RAG in der Praxis – Generierung synthetischer Testdatensätze* untersucht Luka Panic [18] die Testset-Generierung mithilfe von RAGAS. Es treten bei 17 % der generierten Fragen Fehler beim Generieren der Testfragen auf. Dies hat vielfältige Gründe, die von nicht verwertbaren Antworten des LLMs bis zu Verbindungsproblemen oder dem Erreichen des Limits der maximalen Anfragen an APIs reichen.

Auch bei der Bewertung von Antworten können sich ungewollte und bisher noch ungeahnte Biases einschleichen. In dem Paper „Large Language Models are not Fair Evaluators“ [34] wird gezeigt, dass, wenn ein LLM eine von zwei gegebenen Antworten aussuchen müsste, die erste bevorzugt wurde, selbst wenn die gleiche Frage mit einer anderen Reihenfolge gestellt wurde. RAGAS vergleicht keine Antworten miteinander, und daher ist dieser Bias kein direktes Problem für die verwendeten Metriken. Was jedoch einen Einfluss auf die Bewertung von Antworten haben kann, ist der Bias zu gewissen Nummern. Wie in [29]

beschrieben, bevorzugen LLMs bei der Bewertung lieber Zahlen, welche Vielfache von 5 und 10 sind.

Auch die allgemeine stochastische Natur von LLMs spielt eine Rolle, da bei der gleichen Anfrage unterschiedliche Antworten und somit auch Bewertungen zurückgegeben werden. Wie groß diese Abweichungen sind, wird in dieser Arbeit kurz untersucht.

Wie in diesem Paper [10] beschrieben, stellt Gemini 1.5 einen bedeutenden Fortschritt in der multimodalen Verarbeitung großer Kontextfenster dar. Das wirft auch die Frage auf, ob RAGs nicht irrelevant sein könnten und durch LLMs mit großen Kontextfenstern abgelöst werden. Es gibt einige Gründe, die dagegen sprechen: LLMs mit größeren Kontextfenstern werden immer langsamer und teurer; die genauen Kosten sind abzuwarten. Jedes Mal alle Daten in den Kontext zu laden, besonders wenn dies über das Internet geschieht, ist eine weitere Hürde. LLMs fällt es auch schwer, bei zu vielen Informationen noch die relevanten zu finden, was zu schlechteren Antworten führen kann. Diese Faktoren lassen darauf schließen, dass RAGs, die nicht nur eine einfache Suche nutzen, noch länger relevant bleiben.

Inzwischen werden spezielle LLMs wie Pleias-RAG entwickelt, um die Suche mit RAGs zu verbessern. [8]

1.7 Rechtliche Fragestellungen

Am 01.08.2024 trat die Verordnung über Künstliche Intelligenz der Europäischen Union (KI-VO) in Kraft. Die Verordnung setzt Regelungen und Maßstäbe für die Verwendung von KI. RAGs sind gemäß Artikel 3 Nr. 1 KI-VO KI-Systeme und fallen damit in den Anwendungsbereich der KI-VO. Bei der Nutzung oder Bereitstellung von LLMs muss sich an die KI-VO gehalten werden. Die Nutzerinnen der RAGs müssen sich der aus der KI-VO ergebenden Pflichten bewusst sein. Ebenso gilt es, bei Anbindung firmeninterner Dokumente die Datenschutz-Grundverordnung (DS-GVO) zu beachten. Die DS-GVO regelt den Umgang mit personenbezogenen Daten, die in solchen firmeninternen Dokumenten enthalten sein könnten.

2 Methoden und Materialien

2.1 Werkzeuge

Für die RAGs und die Bewertung der RAGs werden Tools benötigt. Im Nachfolgenden werden diese Tools genauer erklärt.

2.1.1 Ollama [16]

Ollama ist ein Open-Source LLM-Server, der auf einem eigenen Computer oder in der Cloud ausgeführt wird. Es können verschiedene Open-Source LLMs und Embedding-Modelle ausgeführt werden. In der vorliegenden Arbeit werden die Modelle ollama/nomic-embed-text und ollama/deepseek-r1:32b verwendet.

2.1.2 Embeddings

Vektoren sind die Beschreibung einer Position im mehrdimensionalen Raum. Es handelt sich hier meist um Positionen, welche mehrere Tausende Dimensionen haben. Embeddings ermöglichen die Darstellung von z. B. Wörtern im mehrdimensionalen Raum. Je näher zwei Wörter im Raum beieinander sind, desto ähnlicher sind sie.

Welche Embeddings für welche Versuche genutzt werden, wird in Kapitel 5.2.3 "Bewertung"(des Kapitels "Konkretisierung der Versuche") beschrieben.

2.1.3 Vektordatenbank ChromaDB [3]

ChromaDB ist eine Open-Source-Vektordatenbank, die zur persistenten Speicherung und effizienten Abfrage von hochdimensionalen Embeddings eingesetzt wird.

Die Vektordatenbank ist also ein fester Bestandteil des RAGs und wird sowohl für die Open-Source-Modelle als auch für die Closed-Source-Modelle von z.B. OpenAI benutzt. Dies schafft eine einheitlichere Basis zum Vergleichen der LLMs.

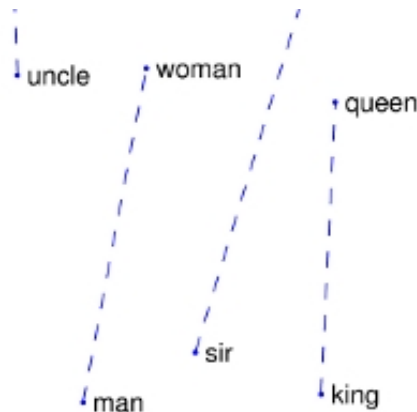


Abbildung 2.1: Beispiel für ein Embedding von "Mann" und "Frau" in einem niedrigerdimensionalen Raum.

2.1.4 Retrieval Augmented Generation Assessment (RAGAS) [7]

RAGAS ist ein Open-Source-Tool und liefert neben dem Tool selbst hilfreiche Dokumentation für die Metriken und die Bewertung von RAGs. Es werden Funktionen wie die automatische Generierung von Interessengruppen, die Testset-Generierung und die Bewertung von RAGs anhand von Testsets bereitgestellt. Für diese Arbeit sind die Funktionen der Testset-Generierung und die damit ermöglichte Bewertung der RAGs relevant.

Was RAGAS von den vorherigen Tools unterscheidet, ist, dass keine „reference answer“ benötigt wird. RAGAS ist beliebt, da es sich gut mit vielen Tools integrieren lässt.

2.1.5 Langchain [13]

Langchain ist ein Framework, welches bei der Entwicklung von Anwendungen, die LLMs nutzen, eine erhebliche Hilfe ist. Durch die vielen Komponenten und Integrationen mit wichtigen Tools bietet es die wichtige Grundlage für die Verbindung von Daten und komplexen Workflows. In dieser Arbeit wird Langchain genutzt, um die Vektordatenbank und Ollama mit RAGAS zu verbinden. Genutzt wurde die Version 0.3.25.

2.2 Daten

Da die Nutzung von RAG-Evaluation-Tools für betriebliche Abläufe untersucht werden soll, werden zum Teil echte, nicht generierte Dokumente – im Folgenden originale Dokumente

genannt – verwendet. Die Dokumente stammen aus Unterlagen eines Einzelunternehmens, welches vereinfachte CMS-Webseiten für Grundschulen entwickelt hat. Die Unternehmung wird nicht mehr aktiv verfolgt, und die anonymisierten Daten können ohne Bedenken für diese Arbeit genutzt werden. In den Versuchen wurden drei unterschiedliche Anzahlen an Dokumenten getestet: 10, 100 und 400. Aus den Unternehmungen des Autors ließen sich 73 nutzbare Dokumente finden. Neben den Dokumenten, welche Businesspläne, Finanzpläne, aber auch Elternbriefe umfassen, gibt es den dazugehörigen Code. Für die zehn Dokumente wurden nur „originale“ Dokumente genutzt. Um von 73 gegebenen Dokumenten auf 100 Dokumente zu kommen, wurden mithilfe von LLMs weitere Dokumente generiert. Beim Generieren der Dokumente wurden dem LLM die bisherigen Dokumente zur Verfügung gestellt und komplett neue Bereiche/Projekte erfunden. Diese bestehen dann aus Kostenplanungen, Zeitplänen und Elternbriefen für Datenschutzinformationen. Für die 400 Dokumente wurde der Code des realen Produktes mit einbezogen. Dieser besteht aus drei Projekten: 1. die Webseite, die öffentlich zugänglich ist, 2. dem Admin-Bereich und 3. dem Backend.

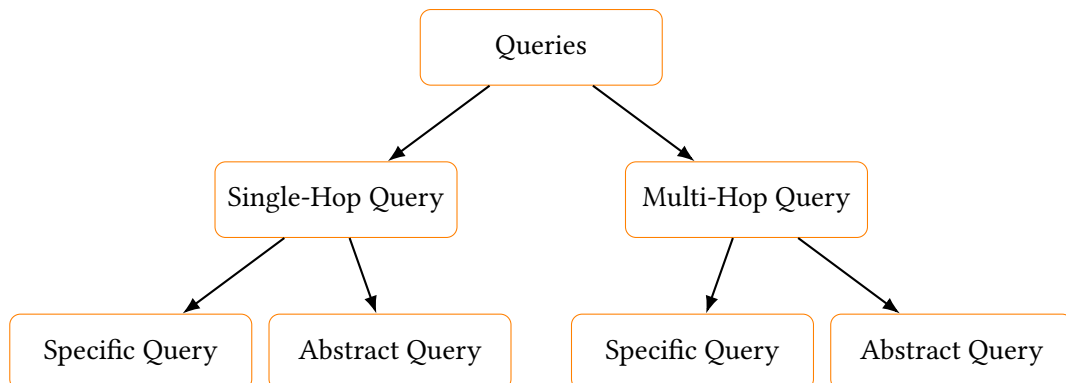
Die folgenden drei Stufen wurden gewählt, um typische Anwendungsszenarien realistisch abzubilden:

1. **10 Dokumente:** Ein einzelner Anwendungsfall – das RAG-System wird nur temporär genutzt und danach verworfen.
2. **100 Dokumente:** Kontinuierliche Nutzung durch eine Einzelperson – das System wächst schrittweise über die Zeit hinweg.
3. **400 Dokumente:** Gemeinsame Nutzung durch mehrere Personen – das RAG muss verschiedene Themenbereiche abdecken und eine breitere Wissensbasis verwalten.

2.3 Fragebögen

2.3.1 Fragetypen

RAGAS unterstützt verschiedene Fragetypen für die Testset-Generierung, die unterschiedliche Aspekte der RAG-Performance evaluieren. Die folgende Abbildung zeigt die verschiedenen Fragetypen, die RAGAS für die Evaluation von RAG-Systemen verwendet:



Diese verschiedenen Fragetypen ermöglichen es, unterschiedliche Aspekte der RAG-Performance zu testen. Während spezifische Fragen häufig mit einer einzigen Anfrage an die Wissensdatenbank beantwortet werden können, benötigen abstrakte Fragen eine Erklärung. In der RAGAS-Dokumentation [27] wird für konkrete Fragen das Beispiel gegeben: „Wann hat Einstein die Relativitätstheorie veröffentlicht?“, während eine abstraktere Frage wäre: „Wie hat Einsteins Relativitätstheorie unser Verständnis der Welt verändert?“

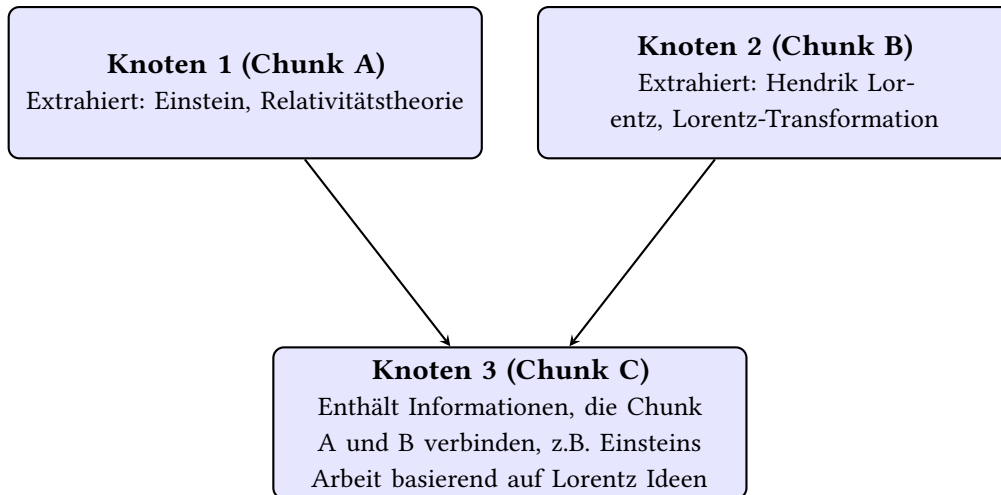
Bei Multi-Hop-Queries handelt es sich um Fragen, die mehr als eine Wissensabfrage benötigen. Die Frage „Welche Wissenschaftler haben Einsteins Relativitätstheorie beeinflusst und welche Theorie haben sie vorgeschlagen?“ benötigt erst eine Abfrage, um die Wissenschaftler herauszufinden, und dann weitere, um die jeweils vorgeschlagene Theorie abzufragen. Für die abstrakte Multi-Hop-Query können wir wieder nach einer Erklärung für den Inhalt und wie sich dieser über die Zeit verändert hat, fragen.

Diese Unterscheidung wird getroffen, um sowohl sehr gezielte Wissensabfragen als auch abstraktere Abfragen über mehrere Dokumente zu testen.

2.3.2 Wissensgraph

Für die eben erwähnten Multi-Hop-Queries müssen aus den gegebenen Dokumenten Themen, welche zusammenhängen, aber nicht direkt im gleichen Dokument vorkommen, gefunden werden. Da dies bei großen Datensätzen manuell oder selbst mit einem LLM schwierig ist, wird ein Wissensgraph erstellt.

Dies geschieht in drei Schritten. Zuerst werden die Dokumente beim sogenannten Chunking in kleinere Einheiten (Knoten) unterteilt. Aus diesen Einheiten können dann Entitäten wie z.B. Namen (Einstein) oder Schlüsselbegriffe (Relativitätstheorie) extrahiert werden. Im letzten Schritt werden dann Verbindungen zwischen Knoten hergestellt.



Für die Daten aus den Versuchen mit 100 Dokumenten hat RAGAS 27 Themen identifiziert. Unter anderem wurden folgende Themen gefunden: Finanzmanagement, Bildungsprojekt Digitalisierung, Projektmanagement und Planung, Zuwendungsverwaltung, Break-Even-Analyse, Finanzplanung und Investitionen, Finanzplanung und Liquidität.

2.3.3 Evaluation

3 Metriken

In diesem Kapitel geht es um die verschiedenen Metriken, die für die Bewertung von RAG-Evaluierungstools verwendet werden können. Metriken sind das Herzstück der Bewertung von RAGs, da sie die Qualität der RAGs bewerten und somit die Entwicklung und den Fortschritt der RAGs messen. Weitere Informationen können in der Dokumentation von RAGAS gefunden werden [24].

In dieser Arbeit werden die vier Metriken Context Precision, Context Recall, Response Relevancy und Faithfulness für die Bewertung genutzt. Die Kombination dieser Metriken deckt die wichtigen Funktionen eines RAGs ab. Sie werden von RAGAS standardmäßig genutzt und wurden zusammen mit der Idee der Fragensgenerierung durch LLMs in dem Paper von RAGAS vorgestellt [5].

3.1 Retrieval Augmented Generation

Diese Metriken basieren auf Faktenextraktion, mithilfe derer sich dann Bewertungen berechnen lassen. Für die Extraktion der Fakten wird häufig ein LLM verwendet, das als Richter fungiert.

3.1.1 Context Precision

Die Kontextpräzision ist eine Metrik, die den Anteil relevanter Textabschnitte in den abgerufenen Kontexten misst. Sie wird als Mittelwert der Präzision@k für jeden Textabschnitt im Kontext berechnet.

$$\text{Kontext-Präzision@K} = \frac{\sum_{k=1}^K (\text{Präzision@k} \times v_k)}{\text{Gesamtzahl der relevanten Elemente in den Top } K \text{ Ergebnissen}}$$

$$\text{Präzision@k} = \frac{\text{richtig positive@k}}{(\text{richtig positive@k} + \text{falsch positive@k})}$$

(eigene Übersetzung nach [21])

Diese Metrik ist für uns als Qualitätskontrolle wichtig, da sie uns sagt, ob es Probleme beim Testen mit dem Vector Store gibt.

Wenn es einen guten Context Precision Score gibt, dann lässt sich hier gut bewerten, ob das LLM in der Lage ist, die relevanten Informationen in dem Kontext zu finden. Da dies ein wichtiger Aspekt eines guten RAGs ist, wird diese Metrik im Rahmen dieser Arbeit betrachtet.

3.1.2 Context Recall

Context Recall misst, wie viele der relevanten Dokumente (oder Informationsstücke) erfolgreich abgerufen wurden. Es konzentriert sich darauf, keine wichtigen Ergebnisse zu verpassen. Ein höherer Recall bedeutet, dass weniger relevante Dokumente ausgelassen wurden. Kurz gesagt geht es beim Recall darum, nichts Wichtiges zu übersehen. (eigene Übersetzung nach [22])

Wenn es einen guten Context Recall Score gibt, dann lässt sich hier gut bewerten, ob das LLM in der Lage ist, die relevanten Informationen in dem Kontext zu finden. Da dies ein wichtiger Aspekt eines guten RAGs ist, wird diese Metrik im Rahmen dieser Arbeit betrachtet.

3.1.3 Response Relevancy

Die Response Relevancy-Metrik misst, wie relevant eine Antwort in Bezug auf die Nutzereingabe ist. Höhere Werte zeigen eine bessere Übereinstimmung mit der Nutzereingabe an, während niedrigere Werte vergeben werden, wenn die Antwort unvollständig ist oder redundante Informationen enthält. (eigene Übersetzung nach [28])

Diese Metrik bildet mit der Noise Sensitivity eine wichtige Grundlage für die Bewertung des RAGs. Denn selbst wenn die Antworten richtig sind, ist die Bewertung des RAGs nicht gut, wenn die Antworten nicht relevant für die Frage sind.

3.1.4 Faithfulness

Die Faithfulness-Metrik misst, wie faktentreu eine Antwort im Vergleich zum abgerufenen Kontext ist.

Eine Antwort gilt als faktentreu, wenn alle ihre Aussagen durch den abgerufenen Kontext gestützt werden können.

Die Berechnung erfolgt nach folgender Formel:

$$\text{Faithfulness Score} = \frac{\text{Anzahl der durch den Kontext gestützten Aussagen in der Antwort}}{\text{Gesamtanzahl der Aussagen in der Antwort}} \quad (3.1)$$

(eigene Übersetzung nach [23])

3.1.5 Context Entities Recall

In diesem Kontext ist eine Entität eine Informationseinheit, die im Kontext vorkommt. Dies könnte z.B. ein Name, ein Ort, ein Datum oder eine andere Informationseinheit sein.

Die Context Entity Recall-Metrik misst den Recall des abgerufenen Kontexts, basierend auf der Anzahl der Entitäten, die sowohl in der Referenz als auch im abgerufenen Kontext vorkommen, relativ zur Gesamtanzahl der Entitäten in der Referenz.

Einfach ausgedrückt misst sie, welcher Anteil der Entitäten aus der Referenz im abgerufenen Kontext wiedergefunden wird.

(eigene Übersetzung nach [20])

Diese Metrik ist als Qualitätskontrolle wichtig, da sie aussagt, ob es Probleme beim Testen mit dem Vector Store gibt.

3.1.6 Noise Sensitivity

Noise Sensitivity misst, wie häufig ein System Fehler macht, indem es falsche Antworten gibt, wenn entweder relevante oder irrelevante abgerufene Dokumente verwendet werden. Um die Noise Sensitivity zu bestimmen, wird jede Aussage in der generierten Antwort daraufhin überprüft, ob sie auf der Grundlage der Referenz korrekt ist und ob sie dem relevanten (oder irrelevanten) abgerufenen Kontext zugeordnet werden kann. (eigene Übersetzung nach [25])

3.1.7 Multimodal Faithfulness/Multimodal Relevance

RAGAS bietet auch Metriken für MLLMs. Da es in dieser Arbeit nur um LLMs geht, sind diese nicht für diese Arbeit relevant.

3.2 Nvidia Metrics

Diese erheben keinen Anspruch auf Objektivität, sie fragen das LLM direkter nach einer Bewertung und nutzen es nicht zur Extraktion von Daten für weitere Berechnungen. Hier werden einzelne Bewertungen generiert, welche keinen tieferen Einblick in die Bewertung gewähren.

3.2.1 Answer Accuracy

Answer Accuracy misst die Übereinstimmung zwischen der Antwort eines Modells und einer Referenz (Ground Truth) für eine gegebene Frage. Dies geschieht über zwei verschiedene LLMs, diese geben jeweils eine Bewertung (0, 2 oder 4) zurück. Die Metrik wandelt diese Bewertungen in eine Skala von $[0,1]$ um und nimmt dann den Durchschnitt der beiden Bewertungen der Richter.

(eigene Übersetzung nach [26])

Diese Metrik nutzt ein LLM. Die Antwort und die Musterlösung werden dem LLM zur Bewertung vorgelegt. Da es hier zu einem positional Bias kommen kann, wird das LLM zweimal nach einer Bewertung gefragt, jeweils mit einer anderen Reihenfolge. Dies hat Vorteile gegenüber der Answer Correctness, da es weniger Aufrufe mit weniger Tokens an das LLM braucht. Es werden im Vergleich zur Answer Correctness auch robustere Bewertungen getroffen, jedoch werden weniger Einblicke in die Bewertung ermöglicht.

3.3 Natural Language Comparison

3.3.1 Factual Correctness

Diese Metriken basieren zum Teil auf der Wahrheitsmatrix (Confusion Matrix), welche die vier Kategorien True Positive, False Positive, False Negative und True Negative definiert [33]. Aus dieser Matrix lassen sich dann Precision, Recall und F1 Score berechnen.

True Positive (TP) = Anzahl der Aussagen in der Antwort, die auch in der Referenz enthalten sind
False Positive (FP) = Anzahl der Aussagen in der Antwort, die nicht in der Referenz enthalten sind
False Negative (FN) = Anzahl der Aussagen in der Referenz, die nicht in der Antwort enthalten sind

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3.2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3.3)$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.4)$$

[33]

3.3.2 Semantic Similarity

Diese Metrik nutzt Embeddings, um zu messen, wie ähnlich die Antwort der Musterlösung ist.

3.4 Non LLM String Similarity

Die String-Ähnlichkeitsmetrik wird ohne LLM berechnet. Das bietet Vorteile hinsichtlich Geschwindigkeit und Kosten im Vergleich zur LLM-Variante.

3.4.1 BLEU Score

Der BLEU Score misst die Ähnlichkeit zwischen der Antwort und der Referenz. Dabei wird die Wortanzahl der Referenz berücksichtigt und eine entsprechende Bestrafung für zu kurze Antworten eingeführt.

3.4.2 ROUGE Score

Mithilfe von n-gram Recall, Precision und dem F1 Score wird die Ähnlichkeit zwischen der Antwort und der Referenz berechnet.

3.4.3 String Presence

Eine einfache Metrik, um zu sehen, ob die Referenz in der Antwort enthalten ist.

3.4.4 Exact Match

Eine noch einfachere Metrik, die nur prüft, ob die Antwort exakt der Referenz entspricht. Diese ist für einzelne Wörter sinnvoll.

3.5 General Purpose

Dies sind Metriken, welche manuell konfiguriert werden müssen, aber eine gute Bewertung der Qualität eines RAGs liefern können. Die Metriken reichen von einfachen Fragen, wie „Ist die Antwort schädlich“ oder „Hat die Intention des Users verletzt“, bis hin zu komplexeren, einleitend definierten Bewertungen.

- Aspect Critic: Dem LLM können eigene Kriterien vorgegeben werden, z.B. „Ist die Antwort schädlich?“; hier gibt es nur Ja oder Nein als Antwort.
- Simple Criteria Scoring: Ähnlich wie der Aspect Critic, jedoch mit einer Zahl als Antwort.
- Rubrics based Scoring: Erlaubt eine Bewertung mit Vorgaben, welche Kriterien für welchen Score erfüllt sein müssen.
- Instance Specific Rubrics Scoring: Ist sehr ähnlich zu Rubrics based Scoring, erlaubt jedoch noch genauere Definition von Bewertungskriterien pro Frage.

3.6 Andere Metriken

3.6.1 Summarization

Summarization ist die Anzahl der richtig beantworteten Fragen geteilt durch die Anzahl aller Fragen. Dies ist eine sehr einfache und oberflächliche Metrik.

4 Ähnliche Arbeiten

4.1 RAG Evaluation: Assessing the Usefulness of Ragas

Das Team von Beatrust hat im Februar 2024 eine Reihe zu RAGs veröffentlicht [1]. Es werden unter anderem die Notwendigkeit und auch die einzelnen Metriken von RAGAS erklärt. Im dritten Artikel dieser Reihe unternehmen sie einen Versuch, die Nützlichkeit von RAGAS zu untersuchen.

Der Versuch besteht aus 50 Fragen aus einem Interessensfeld des Autors. Diese wurden von einem RAG mit GPT-4 und einem mit GPT-3.5-turbo beantwortet und dann sowohl von RAGAS als auch von ihm bewertet.

Der Autor kommt zu dem Ergebnis, dass RAGAS geeignet ist, um RAGs zu bewerten und besser ist, als die Bewertung von Langchain. Es wird jedoch angemerkt, dass der Autor eine höhere Übereinstimmung mit seinen Ergebnissen erwartet hätte.

4.2 Benchmarking Large Language Models in Retrieval-Augmented Generation

In ihrem Paper „Benchmarking Large Language Models in Retrieval-Augmented Generation“ [2] haben Jiawei Chen, Hongyu Lin, Xianpei Han and Le Sun ihre eigenen Metriken entwickelt, um die Fähigkeit von RAGs zu bewerten.

Die vier Metriken sind

- **Noise Robustness (Rauschrobustheit)**, die untersucht das Verhalten, wenn mehr Informationen gegeben sind als notwendig wären, um die Frage zu beantworten. Dies könnte die Frage nach einem bestimmten Ereignis sein und das Rauschen würde ein Dokument zu einem anderen Ereignis sein.
- **Negative Rejection (Negative Ablehnung)**, werden dem LLM nur irrelevante Dokumente zur Verfügung gestellt. Das LLM sollte in diesem Fall antworten, dass es die Frage nicht beantworten kann.
- **Information Integration (Informationsintegration)**, untersucht wie gut ein LLM zwei Fragen in einem aus mehreren Dokumenten beantworten kann.

- **Counterfactual Robustness (Kontrafaktische Robustheit)**, die dem LLM zwei Dokumente mit widersprüchlichen Informationen gibt.

Diese Metriken werden von den Autoren Retrieval-Augmented Generation Benchmark (RGB) genannt. Mithilfe von RGB bewerten sie in englischer und chinesischer Sprache sechs damals state of the art Modelle ChatGPT (OpenAI 2022), ChatGLM-6B (THUDM 2023a), ChatGLM2-6B (THUDM 2023b), Vicuna-7B-v1.3 (Chiang et al. 2023), Qwen-7B-Chat (QwenLM 2023), BELLE-7B-2M (Yunjie Ji 2023)

Es ließ sich zeigen, dass die LLMs generell gut darin sind, einfache Fragen auch bei irrelevanten Dokumenten zu finden. Wenn sich die Information jedoch über einen größeren Text verteilt, dann haben die LLMs Schwierigkeiten. Auch eine hohe Rauschrate von über 80 % führt zu einer deutlichen Verschlechterung.

Bei der „Negative Ablehnung“ war das beste Ergebnis für Englisch bei 45 % und 43,33 % für Chinesisch. Dies bedeutet, dass die LLMs Anweisungen missachten und sich verwirren lassen, wenn es keinen relevanten Inhalt gibt.

Bei der Informationsintegration, also dem Beantworten einer komplexeren Frage über mehrere Dokumente, sank die Genauigkeit der Antworten deutlich. Es wurde ein maximaler Score von 60 % für englische und 67 % für chinesische Fragen erreicht. Wenn auch Rauschen mit hinzugefügt wird, sinkt die Genauigkeit weiter auf 43 % und 55 %.

Die letzte „Metrik Kontrafaktische“ Robustheit zeigte, dass, selbst wenn das LLM selber die richtige Antwort kennt, es den falschen Informationen vertraut. Das ist ein großes Problem für RAGs und deren Zuverlässigkeit!

Insgesamt zeigt das Paper auf, dass LLMs 2023 noch einige Probleme mit wichtigen Bereichen eines RAGs haben.

4.3 RAG-Bewertungsprozess

Das Flussdiagramm veranschaulicht die drei Hauptphasen des RAG-Bewertungsprozesses:

1. Dokumentenverarbeitung

- Dokumente werden geladen und in Abschnitte unterteilt
- Textabschnitte werden eingebettet
- Eingebettete Vektoren werden in ChromaDB gespeichert





2. Erstellung Fragebögen

- Verwendet LLM zur Generierung von Fragen
- Erstellt Testsets mit Fragen und Referenzantworten

3. Bewertungsprozess

- Verwendet das generierte Testset
- Ruft Kontext aus ChromaDB ab
- Bewertet Modellantworten mit LLM als Richter
- Generiert umfassende Bewertungsberichte

Legende für Flussdiagrammfarben:

-  **Modell:** (z.B. LLMs, Einbettungsmodelle)
-  **Speicher:** (z.B. Vektorspeicher, ChromaDB)
-  **Prozess:** (z.B. Dokumentenlader, Bewertung)
-  **Daten:** (z.B. Dokumentensammlung, Testset, Bericht)

Das Diagramm hebt hervor, wie bestimmte Komponenten, wie LLM, für verschiedene Zwecke wiederverwendet werden, während separate Einbettungsmodelle für spezifische Aufgaben beibehalten werden. Dieser modulare Ansatz ermöglicht flexible Versuche mit verschiedenen Modellen und Konfigurationen, während ein konsistentes Bewertungsframework beibehalten wird.

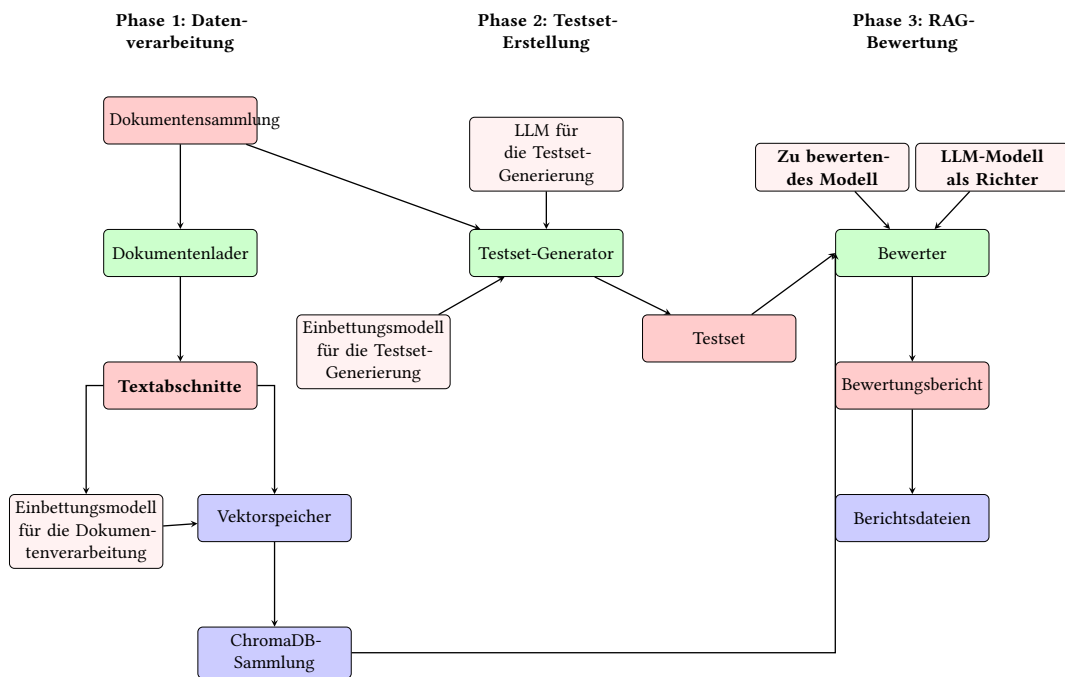


Abbildung 4.1: Flussdiagramm des RAG-Bewertungsprozesses, das die Interaktion zwischen verschiedenen Komponenten und Modellen zeigt. Spezifische Modellnamen (z.B. gpt-4-turbo, text-embedding-3-large) sind im Haupttext beschrieben.

5 Versuche

5.1 Versuchsplan

Um systematisch zu bewerten, ob RAG-Bewertungstools für den Einsatz in kleinen und mittleren Unternehmen (KMU)¹ bereit sind, sind umfassende Versuche erforderlich. Der folgende Versuchsplan skizziert die wichtigsten Variablen, die Methodik und die Bewertungskriterien.

5.1.1 Forschungsfragen

Der Versuch wird die folgenden zentralen Forschungsfragen behandeln:

1. Sind aktuelle RAG-Bewertungsframeworks in Bezug auf Kosten, Komplexität und Ressourcenanforderungen für den Einsatz in einem KMU geeignet?
2. Wie beeinflussen verschiedene Dokumententypen und Datenvolumina die Qualität von Abruf und Generierung?
3. Wie zuverlässig und konsistent sind die verfügbaren Bewertungsmetriken zur Beurteilung der RAG-Leistung?
4. Was ist das optimale Gleichgewicht zwischen Kosten, Leistung und Implementierungskomplexität für jeden Anwendungsfall in einem KMU?

5.1.2 Variablen in den Versuchen

Dokumententypen

Verschiedene Dokumentenformate werden getestet, um die Vielseitigkeit des Systems zu bewerten:

- PDF (.pdf)
- Klartext (.txt)

¹6.

- Word-Dokumente (.docx, .doc)
- Excel-Tabellen (.xlsx, .xls)
- CSV-Dateien (.csv)
- E-Mails (.eml)
- PowerPoint-Präsentationen (.pptx, .ppt)

Datenvolumen

Die Skalierbarkeit des Systems wird, wie bereits beschrieben, mit unterschiedlichen Datenmengen getestet: 10, 100 und 400 Dokumente.

- Für die Versuche mit **10 Dokumenten** werden existierende Dokumente ausgewählt.
- Für die Versuche mit **100 Dokumenten** müssen zusätzliche Dokumente generiert werden, vorzugsweise mit einem LLM.
- Für die Versuche mit **400 Dokumenten** wird zusätzlich Code verwendet.

Modelle zur Bewertung

Mehrere Modelle werden bewertet, die verschiedene Kostenschichten und Fähigkeiten repräsentieren. Hierbei ist es wichtig zu überlegen, welche Optionen für KMU gültige Anwendungsfälle sind. **Open-Source-Modelle** (z.B. Llama 2, Mistral 7B, Deepseek R1) bieten eine Vielzahl von Vorteilen, wie die Möglichkeit, sie zu modifizieren und mehr Kontrolle über die Daten zu haben. Entscheidend ist zudem die technische Kompetenz, welche benötigt wird, um diese Modelle selbst zu hosten. **Mittelklasse-API-Modelle** (z.B. Claude Haiku, GPT-3.5 Turbo) sind günstiger als die Hochleistungsmodelle und bieten dennoch eine gute Leistung. Da sie nicht Open Source sind, bieten sie weniger Kontrolle über die Daten und das Modell selbst. Manchmal muss man mehr für private Instanzen zahlen. **Hochleistungsmodelle** (z.B. GPT-4, Claude 3 Opus) sind die teuerste Option, bieten aber auch die beste Leistung, sowohl in Bezug auf Geschwindigkeit als auch auf die Qualität der generierten Antworten.

Bewertungsmetriken

Während des Versuchs wird neben der menschlichen Bewertung RAGAS zur Bewertung verwendet. RAGAS wird die beschriebenen Metriken generieren, die später verglichen und bewertet werden können. Die menschliche Bewertung wird als subjektives Maß verwendet, um die Ergebnisse von RAGAS zu vergleichen.

5.1.3 Kosten- und Zeitanalyse

Die Kosten für die Bewertung eines RAGS werden mithilfe der in RAGAS eingebauten Funktionen berechnet. Die Zeit, welche die Ausführung braucht, wird ebenfalls für die Bewertungen gemessen.

5.1.4 Versuchsprotokoll

1. **Dokumentensammlung und -vorbereitung** Die Dokumente werden in allen oben genannten Zielformaten gesammelt.
2. **Testset-Generierung** Verschiedene Fragetypen (faktisch, inferentiell, vergleichend) werden generiert und Referenzantworten zur Bewertung erstellt. Dies geschieht automatisch durch das RAGAS-Framework. Das Testset wird manuell auf Qualität und Abdeckung validiert, wobei dies anhand einer Reihe zufälliger Proben erfolgt.
3. **Systemkonfiguration** Die Einbettungsmodelle und Parameter werden konfiguriert, Vektorspeicher mit konsistenten Einstellungen eingerichtet und die Bewertungsframeworks implementiert.
4. **Durchführung der Bewertung** Die hochgeladenen Dateien, generierten Dokumente und das Testset werden wiederverwendet. Im ersten Schritt werden diese Daten erstellt. Anschließend wird die Bewertungspipeline ausgeführt und die Ergebnisse werden aufgezeichnet.
5. **Analyse und Berichterstattung** Eine vergleichende Analyse über alle Variablen hinweg wird durchgeführt, einschließlich einer Kosten-Nutzen-Analyse für die geschäftliche Entscheidungsfindung und Empfehlungen für optimale Konfigurationen.

5.1.5 Bewertungskriterien für die Geschäftstauglichkeit

Die endgültige Bewertung wird RAG-Systeme in diesen Dimensionen bewerten:

- **Implementierungskomplexität:** Wie schwierig ist die Einrichtung und Wartung?
- **Kostenvorhersehbarkeit:** Sind die Kosten stabil und vorhersehbar?
- **Leistungszuverlässigkeit:** Sind die Ergebnisse konsistent und nicht komplett anders bei jeder Bewertung?
- **Skalierbarkeit:** Wie gut bewältigt das System wachsende Datenanforderungen?

Dieser Ansatz mit Versuchen bietet einen umfassenden Rahmen, um zu bewerten, ob aktuelle RAG-Bewertungstools ausreichend ausgereift für die Einführung in einem KMU sind, mit klaren Anleitungen zu optimalen Konfigurationen und Implementierungsstrategien.

5.2 Konkretisierung der Versuche

5.2.1 Dokumentenverarbeitung

Damit die Dokumente in der Vektordatenbank gesichert werden können, müssen sie erst in Vektoren konvertiert werden. Hier benutzt der Autor Embeddings von OpenAI sowie Open-Source-Embeddings von nomic.ai [15].

Embedding-Modell	10	100	400
openai/text-embedding-3-large	X	X	X
ollama/nomic-embed-text	X	X	X

Tabelle 5.1: Kombinationen aus Dokumentenanzahl und Embedding-Modell für die Versuche (X = Kombination wird getestet)

5.2.2 Testset-Generierung

Um die optimale Anzahl an Fragen pro Testset zu untersuchen, werden folgende Kombinationen generiert:

Dokumentenanzahl	Anzahl Fragen pro Testset	Anzahl Testsets pro Modell
10	15, 30	2
100	50, 100	2
400	150, 300	2
Summe Testsets pro Modell		6

Tabelle 5.2: Kombinationen aus Dokumentenanzahl und Testset-Größe

5.2.3 Bewertung

Um die Robustheit und Übertragbarkeit der Bewertungsergebnisse zu erhöhen, werden alle Kombinationen aus Embedding-Modell und Bewertungsmodell getestet. Das bedeutet, dass für jedes Testset sowohl **openai/text-embedding-3-large** als auch **ollama/nomic-embed-text** als Embedding-Modell verwendet werden und die Bewertung jeweils mit **GPT-4** sowie **Deepseek-R1 (ollama/deepseek-r1:7b)** erfolgt. Insgesamt ergeben sich so 24 Versuche (2 Embeddings \times 2 Bewerter \times 6 Testset-Varianten).

Verwendete Abkürzungen in der Tabelle:

- OAI-E = **openai/text-embedding-3-large**

Versuch	Embedding	Dokumente	Fragen	Bewerter	Richter	Wdh.
1	OAI-E	10	15	GPT-4	GPT-4	1
2	OAI-E	10	30	GPT-4	GPT-4	4
3	OAI-E	100	50	GPT-4	GPT-4	1
4	OAI-E	100	100	GPT-4	GPT-4	4
5	OAI-E	400	150	GPT-4	GPT-4	1
6	OAI-E	400	300	GPT-4	GPT-4	1
7	OLL-E	10	15	DSK-R	DSK-R	1
8	OLL-E	10	30	DSK-R	DSK-R	1
9	OLL-E	100	50	DSK-R	DSK-R	1
10	OLL-E	100	100	DSK-R	DSK-R	1
11	OLL-E	400	150	DSK-R	DSK-R	1
12	OLL-E	400	300	DSK-R	DSK-R	1

Tabelle 5.3: Übersicht aller 24 zu generierenden Bewertungsberichte mit Abkürzungen und Wiederholungen

- **OLL-E** = **ollama/nomic-embed-text**
- **GPT-4** = **openai/gpt-4**
- **DSK-R** = **ollama/deepseek-r1:7b**

6 Ergebnisse und Diskussionen

6.1 Ergebnisse aus den Versuchen

6.1.1 Generierte Fragebögen

Da insgesamt 1.290 Fragen generiert wurden, lassen sich diese aufgrund des zeitlichen Aufwandes nicht alle bewerten. Aus jedem der sechs generierten Fragebögen werden stichprobenartig 10 Fragen ausgesucht und überprüft, wie sinnvoll diese sind.

Deepseek/Nomic

Bei der Generierung von Fragen mit DeepSeek kam es zu mehreren Problemen. Bei dem Fragenset, welches 300 Fragen umfassen sollte, traten folgende Probleme auf:

- Von den angeforderten 300 Fragen wurden nur 267 Fragen (89 %) überhaupt generiert; der Rest wurde aufgrund von technischen Problemen oder ungültigen Antworten seitens DeepSeek nicht generiert.
- Von diesen sind 101 zu Themen rund um Bezahlmethoden, Versand und Ähnlichem. In den Dokumenten, welche DeepSeek zur Verfügung gestellt wurden, traten diese Themen nicht auf. Diese Fragen sind daher als ungültig bewertet worden.

Am Ende bleiben 166 von 300 Fragen übrig. **Ca. 45 %** der angeforderten Fragen sind irrelevant!

Beim Generieren des Testsets mit 100 Fragen zeigte sich eine leichte Verbesserung:

- Von den 100 angeforderten Fragen wurden 88 Fragen generiert. Ganze 12 % wurden hier nicht generiert.
- Dieses Mal sind jedoch nur 4 Fragen zu irrelevanten Themen wie Bezahlmethoden, Versand etc.

Am Ende hat das Testset mit 100 Fragen eine **Fehlerquote** von **16 %**.

Aus der Tabelle 5.1 wird ersichtlich, dass die **Fehlerquote** in den Testsets für die 400 Dokumente deutlich größer ist. Der Grund dafür wird noch untersucht.

Angefragt	Generiert	Irrelevant	Fehlerquote
15	11	0	27 %
30	27	7	33 %
50	40	0	20 %
100	88	4	16 %
150	137	49	41 %
300	267	101	45 %

Tabelle 6.1: Übersicht der generierten Fragen und Fehlerquoten pro Testset für DeepSeek

OpenAI

Angefragt	Generiert	Irrelevant	Fehlerquote
15	12	0	20%
30	30	1	3%
50	48	0	4%
100	95	1	6%
150	150	8	5%
300	300	16	5%

Tabelle 6.2: Übersicht der generierten Fragen und Fehlerquoten pro Testset für OpenAI

—

6.1.2 Manuelle Auswertung der Fragebögen

DeepSeek

Bei der manuellen Sichtung der Testsets wurden weitere Fehler entdeckt.

Beispiel 1:

Neben dem vorhin angesprochenen Problem mit den irrelevanten Themen hat DeepSeek auch zwischendurch Fragen und beispielhafte Antworten auf Englisch generiert. Frage: „How much does it cost?“

Antwort: „For orders under \$50, shipping costs \$5.99.“

Diese Frage ist auf Englisch, obwohl explizit Deutsch als Sprache angegeben wurde. Zudem ist diese Frage nicht sinnvoll, da es nie um Lieferungen oder Lieferkosten in den Daten ging.

Beispiel 2:

Bei dieser Frage hat das LLM verdreht, wer bezahlt, und fragt, wie viel die Schulen weniger **verdienen** und nicht, wie viel sie weniger **bezahlen**. Develop 4 Future wird von den Schulen bezahlt und bietet das Produkt den Pilotenschulen zu einem günstigeren Preis an, dadurch entstehen die niedrigeren Einnahmen seitens Develop 4 Future, nicht seitens der Grundschulen.

Frage:

„Hallo! Ich bin Schulleiter/in und überlege, ob wir als Pilotenschule bei Develop 4 Future teilnehmen sollen. Könnt ihr mir sagen, wie viel weniger die beiden Pilotenschulen im ersten Jahr verdienen verglichen mit anderen Schulen?“

Antwort:

„Die beiden Pilotenschulen verdienen im ersten Jahr 2.000 € weniger als die anderen Schulen.“

Diese Frage ist also, auch wenn sie erst gut aussieht, inhaltlich falsch.

Beispiel 3:

Auch gab es Probleme mit Fragen, die zu allgemein gefasst waren.

„Ich möchte wissen, wie die Verfügbarkeit der Webseite für Schulen ist.“

Hier ist nicht geklärt, worauf sich die Verfügbarkeit bezieht. Es könnte sich hier sowohl um die Frage handeln, ob aktuell eine Webseite gekauft werden kann, als auch, wie viel Prozent Erreichbarkeit garantiert wird.

Beispiel 4:

Ebenso ist „Wie hoch ist die Gesamtsumme der Passiva?“ eine Frage, welche nicht spezifiziert, um welches Jahr es sich handelt und ist daher fehleranfällig.

Beispiel 5:

Es gab auch Fragen, die aufgrund des Kontextes verwirrend waren.

Frage:

„Hallo, ich bin ein kanadischer Student, der sich für die Schulsysteme in Deutschland interessiert. Könntest du mir erklären, warum sich die meisten Grundschulen in NRW befinden?“

Antwort:

„Die meisten Grundschulen befinden sich in NRW, damit sie das vom Bundesland zur Verfügung gestellte System Logineo einbinden können, das Lehrer- und Schülerverwaltung bietet.“

Dies ist nicht richtig, die Anzahl der Grundschulen richtet sich nach der demografischen Verteilung der Bevölkerung.

Testset	Fehlerhafte Fragen
Ollama – 10 Dok (15 Fragen)	2
Ollama – 10 Dok (30 Fragen)	6
Summe 10 Dok: 8 / 20 = 40%	
Ollama – 100 Dok (50 Fragen)	2
Ollama – 100 Dok (100 Fragen)	4
Summe 100 Dok: 6 / 20 = 30%	
Ollama – 400 Dok (150 Fragen)	5
Ollama – 400 Dok (300 Fragen)	8
Summe 400 Dok: 13 / 20 = 65%	
Gesamt (Ollama): 27 / 60 = 45%	

Tabelle 6.3: Anzahl fehlerhafter Fragen pro Testset und Gesamtübersicht für Ollama

Die Fehlerquote von 45 % zeigt, dass die Fragen, die DeepSeek generiert, nicht einfach eingesetzt werden können. Es wird deutlich, dass diese Fragen weit entfernt davon sind, die Qualität von menschlich erstellten Fragen zu erreichen.

OpenAI

Bei ChatGPT wurden auch Mängel bei der manuellen Überprüfung festgestellt.

Beispiel 1:

Die Fragen werden so gestellt, dass sie den „gegebenen Kontext“ bewerten sollen. Es fehlen dadurch wichtige Informationen, welche zum Finden der relevanten Dokumente notwendig sind. „Analysieren Sie den bereitgestellten Kontext und erläutern Sie unter der Voraussetzung, dass Sie keine externen Quellen verwenden dürfen, welches zentrale Thema oder welcher Hauptzweck in dem Textabschnitt behandelt wird. Begründen Sie Ihre Antwort anhand spezifischer Textstellen.“

Beispiel 2:

Bei einer Frage war das vorliegende Dokument ein Fragebogen. Fehlerhafterweise wurde die erste Option als die richtige Antwort verstanden, da der Fragebogen nicht ausgefüllt ist, ergibt dies keinen Sinn.

Beispiel 3:

Auch eine Frage, welche die Antwort schon beinhaltet, wurde generiert: „Kannst du mir erklären, was das besondere Merkmal des neuen Schulwebseiten-Systems ist, das ich als Lehrer verwenden werde, um Abwesenheitsmeldungen schnell und einfach zu veröffentlichen?“

Testset	Fehlerhafte Fragen
OpenAI – 10 Dok (15 Fragen)	0
OpenAI – 10 Dok (30 Fragen)	3
Summe 10 Dok: 3 / 20 = 15%	
OpenAI – 100 Dok (50 Fragen)	3
OpenAI – 100 Dok (100 Fragen)	1
Summe 100 Dok: 4 / 20 = 20%	
OpenAI – 400 Dok (150 Fragen)	5
OpenAI – 400 Dok (300 Fragen)	7
Summe 400 Dok: 12 / 20 = 60%	
Gesamt (OpenAI): 19 / 60 = 31.67%	

Tabelle 6.4: Anzahl fehlerhafter Fragen pro Testset und Gesamtübersicht für OpenAI

Wenn wir die Testsets mit Code (150/300 Fragen) ignorieren, kommen wir auf eine **Fehlerquote** von 17.5 %. Dies ist die Hälfte von Ollamas 35 %, also eine deutliche Verbesserung, jedoch immer noch eine beachtliche Menge!

Dennoch offenbaren sich im Vergleich zu einem von Menschen erstellten Fragebogen deutliche qualitative Unterschiede bezüglich der Fragenformulierung.

6.1.3 Auswertung der Reports

Für die Auswertung der Reports werden wieder dieselben Fragen wie vorher aus den Testsets verwendet. Dabei wird geprüft:

- Ist die Frage an sich richtig? Das heißt, ergibt es Sinn, mit dem ursprünglich gegebenen Kontext diese Frage zu stellen?
- Wurde die Frage vom RAG richtig beantwortet?
- Ist die Bewertung der vier Metriken richtig?
- Auffällig war, dass Answer Relevancy am häufigsten abweichend war, deswegen wurde hier zusätzlich bewertet, ob die Bewertung besser oder schlechter sein sollte.

Manuelle Auswertung DeepSeek

Bei der manuellen Bewertung fällt auf, dass ganze 64 % nicht richtig beantwortet wurden. Dabei muss jedoch beachtet werden, dass 43 % der Fragen erst gar nicht sinnvoll sind.

Auch die nicht bewerteten Metriken sind mit bis zu 48 % zu einem großen Teil unbrauchbar. Dies liegt wieder daran, dass das LLM zu lange zum Antworten braucht oder eine ungültige Antwort geliefert hat.

Metrik	Richtig	Falsch	Nicht bewertet
Richtige Frage	31 (51.7%)	29 (48.3%)	–
Gültige Antwort	22 (36.7%)	38 (63.3%)	–
context_precision	39 (65.0%)	1 (1.7%)	20 (33.3%)
faithfulness	29 (48.3%)	2 (3.3%)	29 (48.3%)
context_recall	60 (100.0%)	–	–
answer_relevancy	44 (73.3%)	15 (25.0%)	1 (1.7%)
answer_relevancy sollte höher sein	4 (100.0%)	–	–

Tabelle 6.5: Verteilung der Bewertungen für DeepSeek (mit Prozentangaben)

Manuelle Auswertung OpenAI

Bei der Nutzung der OpenAI API für GPT-4 kam es zu keinen Timeouts oder Ähnlichem, welche zu ungültigen Werten führen würden. Es kam jedoch zu zwischenzeitlichen Rate Limits. Diese könnten von einer Firma jedoch bei einem Vertragsschluss mit OpenAI erhöht werden.

Metrik	Richtig	Falsch
Richtige Frage	43 (72.9%)	16 (27.1%)
Gültige Antwort	42 (71.2%)	17 (28.8%)
context_precision	56 (94.9%)	3 (5.1%)
faithfulness	51 (86.4%)	8 (13.6%)
context_recall	57 (96.6%)	2 (3.4%)
answer_relevancy	43 (72.9%)	16 (27.1%)
answer_relevancy sollte höher sein	7 (53.8%)	6 (46.2%)

Tabelle 6.6: Verteilung der Bewertungen für OpenAI mit Prozentangaben

GPT-4 schneidet deutlich besser als DeepSeek ab, 27 % an nicht sinnvollen Fragen ist jedoch immer noch ein hoher Wert! Die Hälfte der ungültigen Antworten ist durch sinnlose Fragen bedingt; hier ziehen sich also die schlecht generierten Fragen durch.

Probleme mit Code

Da sowohl bei der Testset-Generierung als auch bei der Bewertung der Testsets, die 400 Dokumente nutzten, höhere **Fehlerquoten** zu beobachten sind, wird dies genauer untersucht.

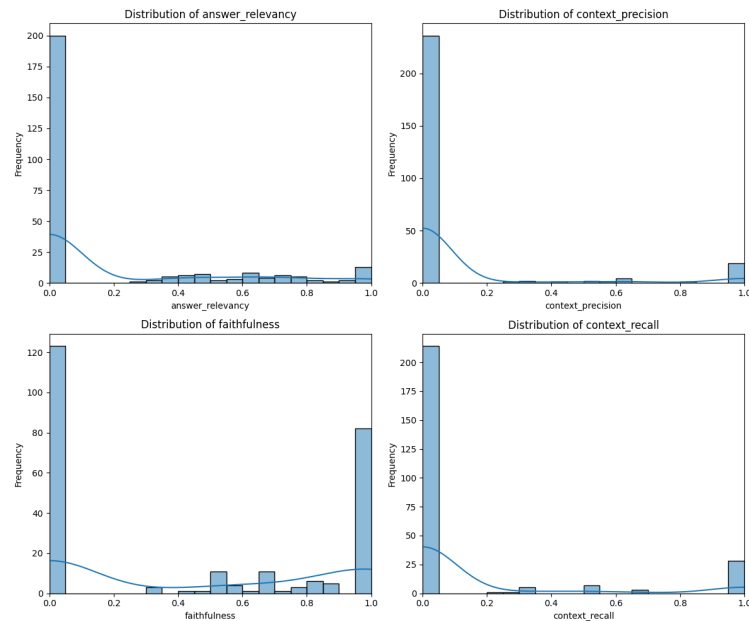


Abbildung 6.1: DeepSeek Ergebnis für 300 Fragen (mit Code-Dokumenten)

Der Vergleich der Anzahl der 0.0-Bewertungen mit DeepSeek (Abbildung 6.1) im Vergleich zu OpenAI (Abbildung 6.2) ist eindeutig.

106 (40 %) der insgesamt 276 zu bewertenden Fragen waren mit komplett 0.0 bewertet worden. Bei der Analyse der speziellen Zeichen im Kontext fällt auf, dass 89 Bewertungen (83 %) zu mehr als 5 % nur aus diesen bestehen. Dies deutet darauf hin, dass DeepSeek starke Probleme hat, Fragen mit Code zu generieren und/oder zu finden.

Ein großer Teil der Fragen hat sich also auf Dokumente mit minderwertiger Qualität bezogen. Durch diese minderwertigen Dokumente hat sich das LLM dann verwirren lassen. Es zeigt sich wieder einmal, dass die Qualität der Daten eine entscheidende Rolle spielt! Bei Betrachtung der Ergebnisse ohne Dokumente, welche Code enthalten, sehen wir, dass die 0.0-Bewertungen bei DeepSeek deutlich zurückgehen. Wie zu erwarten war, ist OpenAI's ChatGPT-4 immer noch deutlich besser.

—

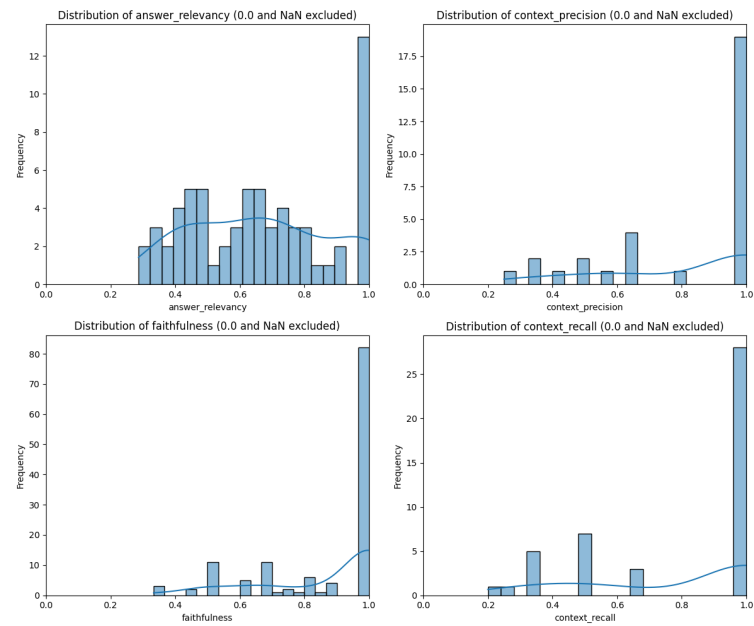


Abbildung 6.2: ChatGPT Ergebnis für 300 Fragen (mit Code-Dokumenten)

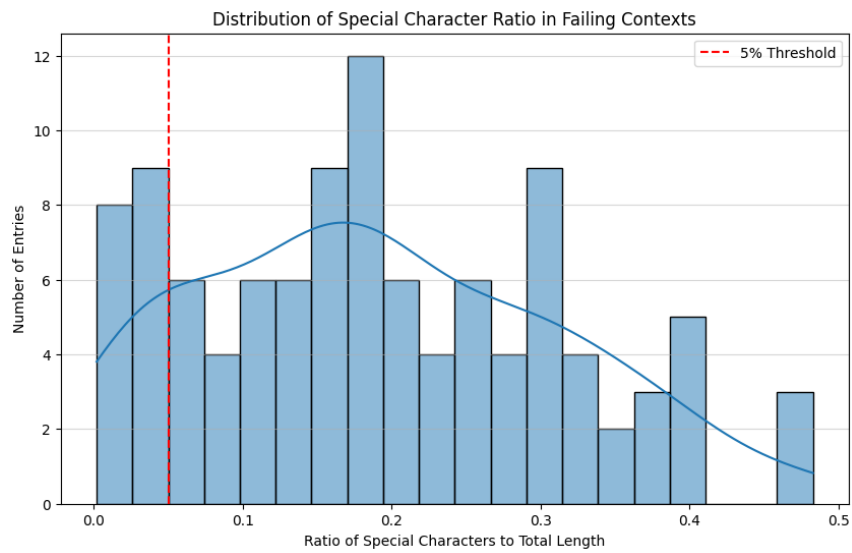


Abbildung 6.3: Abweichungen des Faithfulness Scores bei Code-Dokumenten.

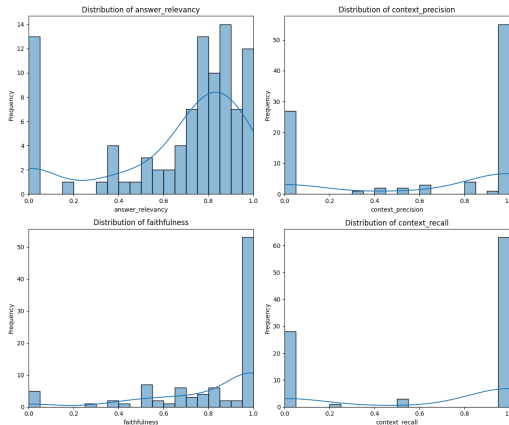


Abbildung 6.4: DeepSeek Ergebnis für 100 Fragen (ohne Code-Dokumente)

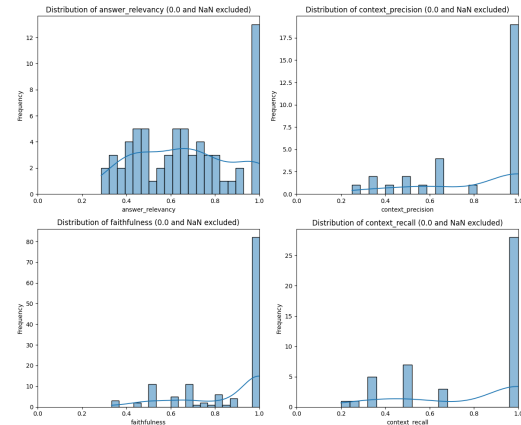


Abbildung 6.5: ChatGPT Ergebnis für 100 Fragen (ohne Code-Dokumente)

6.1.4 Unterschiede über mehrere Durchläufe

Um zu prüfen, wie sich die Ergebnisse von Durchlauf zu Durchlauf unterscheiden, wurden für das Testset mit 100 Fragen für 100 Dokumente mit beiden Modellen vier Durchläufe vorgenommen. In diesem Versuch geht es dann um die Unterschiede pro Durchlauf für das Modell festzustellen und nicht die Modelle miteinander zu vergleichen.

Beim Betrachten der Strip Plots lässt sich gut sehen, dass die Verteilung der Werte pro Durchlauf sehr ähnlich ist und keine großen Abweichungen erkennbar sind. Der Mean und Std wurde über alle Fragen hinweg berechnet.

Metrik	Mean 1	Mean 2	Mean 3	Mean 4	Std 1	Std 2	Std 3	Std 4
Answer Relevancy	0.336	0.366	0.329	0.358	0.346	0.340	0.347	0.361
Faithfulness	0.624	0.593	0.627	0.623	0.442	0.429	0.436	0.453
Context Precision	0.344	0.344	0.344	0.344	0.449	0.449	0.449	0.449
Context Recall	0.415	0.415	0.415	0.415	0.484	0.484	0.484	0.484

Tabelle 6.7: Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für DeepSeek

Beim Betrachten des Durchschnitts und der Standardabweichung lässt sich für die Answer Relevancy und die Faithfulness sehen, dass eine gewisse Schwankung vorhanden ist. Die Metriken für den Kontext sind jedoch sehr konstant!

Bei der Answer Relevancy lässt sich ein Unterschied von 3.7 % feststellen, bei der Faithfulness



Abbildung 6.6: Bewertung der vier Durchläufe mit DeepSeek

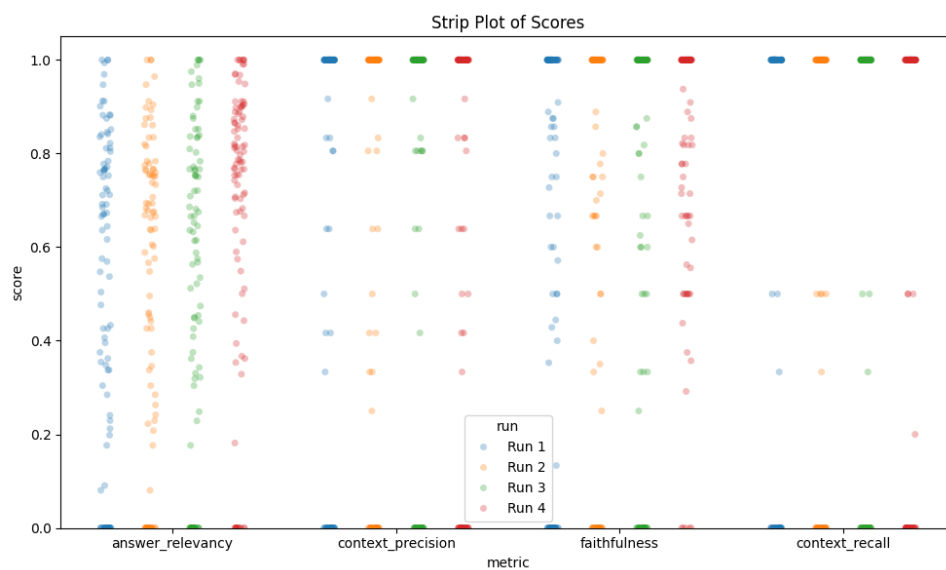


Abbildung 6.7: Bewertung der vier Durchläufe mit GPT-4

3.1 %. Dies liegt für vier Durchläufe im Rahmen, bedeutet aber auch, dass dies beim Einrichten einer automatischen Pipeline berücksichtigt werden sollte.

Metrik	Mean 1	Mean 2	Mean 3	Mean 4	Std 1	Std 2	Std 3	Std 4
Answer Relevancy	0.680	0.681	0.666	0.667	0.320	0.314	0.315	0.313
Context Precision	0.671	0.670	0.666	0.667	0.443	0.445	0.444	0.445
Context Recall	0.647	0.671	0.681	0.665	0.472	0.456	0.458	0.461
Faithfulness	0.846	0.800	0.815	0.823	0.247	0.295	0.273	0.295

Tabelle 6.8: Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für GPT-4

Beim Betrachten des Durchschnitts und der Standardabweichung zeigt sich, dass GPT-4 insgesamt eine konstante Leistung liefert. Die Werte für Answer Relevancy bewegen sich zwischen 66.6 % und 68.1 %, was einer Abweichung von lediglich 1.5 % entspricht. Auch die Standardabweichung ist mit etwa 0,31 gleichmäßig verteilt.

Bei der Faithfulness sind die Mittelwerte etwas variabler und reichen von 80.0 % bis 84.6 %. Dies ergibt eine Differenz von 4.6 %, die im Rahmen liegt, aber auf gewisse inhaltliche Schwankungen hinweist.

Die Metriken für den Kontext (Context Precision und Context Recall) zeigen ebenfalls eine stabile Entwicklung mit geringen Unterschieden in den Mittelwerten. Die Standardabweichungen sind bei diesen Metriken nahezu konstant.

Insgesamt zeigt GPT-4 verlässliche Ergebnisse mit nur geringen Varianzen über mehrere Durchläufe hinweg.

6.1.5 Zuverlässigkeit von Metriken

Um genauer zu untersuchen, wie sich die Metriken bei mehrfacher Ausführung verhalten, wurden die vier Metriken jeweils 50 Mal ausgeführt. Bei der Bewertung wurde immer GPT-4.1 verwendet.

Context Precision & Recall

Beide Metriken wurden 50 Mal bewertet und haben sich wie bei DeepSeek in der Gesamtbewertung als sehr stabil herausgestellt. Es gab in keinem der Durchläufe eine Abweichung.

Answer Relevancy

Hier wurden minimale Abweichungen festgestellt; diese belaufen sich aber auf die zweite Nachkommastelle in der Prozentangabe und sind daher vernachlässigbar.

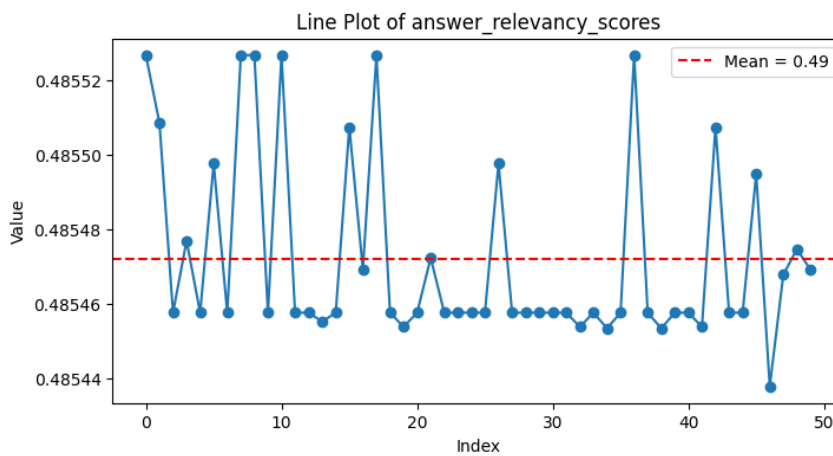


Abbildung 6.8: Abweichungen des Answer Relevancy Scores.

Faithfulness

Bei der Faithfulness sieht dies schon etwas anders aus. Die richtige Bewertung wäre 62.5 %. In 66 % der Fälle war dem auch so, es ist jedoch ersichtlich, dass der Wert teilweise bis zu 12.5 % abweichen kann.

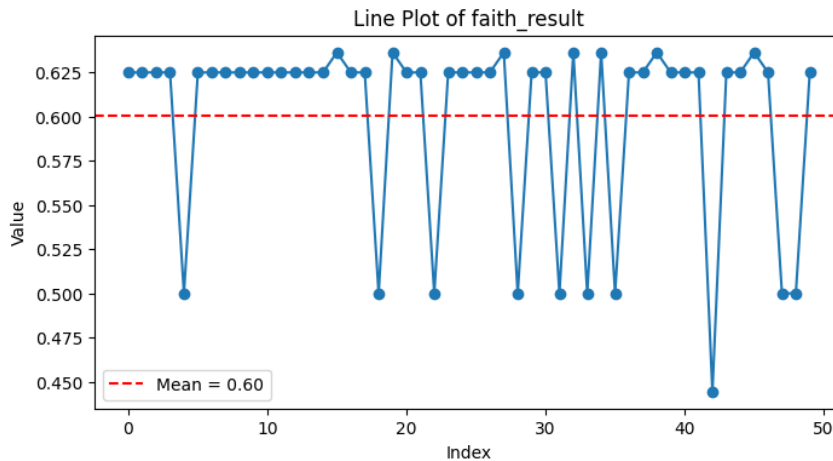


Abbildung 6.9: Abweichungen des Faithfulness Scores.

Die Faithfulness-Metrik hat die größten Schwankungen; dies war auch schon bei dem Versuch mit mehreren Durchläufen ersichtlich.

Ausführungszeiten DeepSeek

Da es bei OpenAI zu den Rate Limits kommen kann, wurde die Anzahl an gleichzeitigen Abfragen von 16 auf 1 reduziert, da es sonst besonders bei längeren Durchläufen zu Problemen kommt. Das führt zu einer deutlichen Verschlechterung der Ausführungszeit. Das Generieren des Testsets mit 15 Fragen dauerte bei 16 gleichzeitigen Anfragen 2 Minuten, bei nur einer maximalen Anfrage wurden es 7 Minuten. Mit diesen Zahlen lässt sich annehmen, dass der Versuch mit 300 Fragen ohne Rate Limit seitens OpenAI sicherlich unter einer Stunde geschafft werden könnte.

Anzahl	Dauer (hh:mm)
15	00:02
30	00:03
50	00:04
100	00:07
150	01:37
300	02:30

Tabelle 6.9: Dauer der Evaluation pro Dokumentenanzahl mit DeepSeek

Ausführungszeiten OpenAI

Für die Bewertung des Testsets mit 300 Fragen (400 Dokumente) wurde Tracing genutzt. Dies lässt uns genauer prüfen, warum gewisse Bewertungen fehlgeschlagen sind. Es kam insgesamt zu 20 Fehlern: 12 Zeitüberschreitungen, weil das LLM nicht innerhalb von 10 Minuten geantwortet hat, acht Antworten waren in einem ungültigen Format, sieben davon für context_recall und eine für faithfulness. Mit den 20 fehlgeschlagenen Metriken kommen wir auf eine **Fehlerquote** von 1.9 %.

Anzahl	Dauer (hh:mm)
15	00:41
30	01:03
50	01:35
100	03:41
150	05:20
300	17:29

Tabelle 6.10: Dauer der Evaluation pro Dokumentenanzahl mit OpenAI

6.1.6 Kostenberechnung

Die Bewertung des RAGs mit den 300 Fragen und OpenAI hat 2 Stunden und 30 Minuten gedauert, dabei sind Kosten in Höhe von 12 Euro entstanden.

Dies kann man mit einer Bewertung, wie in den Versuchen, auf einem Mac Studio (M2 Ultra) vergleichen.

- Laufzeit pro Bewertung DeepSeek: 17 h
- Stromkosten: $0,12 \text{ €/h} \Rightarrow 17 \text{ h} \times 0,12 \text{ €/h} = \mathbf{2,04 \text{ €}}$
- OpenAI API-Kosten pro Bewertung: 12,00 €
Davon sollen 2,00 € lokal durch eigene Ausführung ersetzt werden \Rightarrow verbleibende Abschreibung: **10,00 €/Run**
- Geräteanschaffung: 7.200 € \Rightarrow amortisiert über 720 Runs à 10,00 €
- Gesamtkosten pro Run: $2,04 \text{ € (Strom)} + 10,00 \text{ € (Abschreibung)} = \mathbf{12,04 \text{ €}}$
- Gesamtlaufzeit (720 Runs): $720 \times 17 \text{ h} = 12.240 \text{ h} \approx \mathbf{1 \text{ Jahr, 4 Monate, 10 Tage}}$

Die Preise für die Nutzung eines LLMs über eine Schnittstelle zu entscheiden liegt hier komplett beim Anbieter. Auch die Preise für Strom oder Hochleistungs-GPUs können in Zukunft schwanken. Sollte es also in Zukunft Änderungen an diesen variablen Preisen geben, könnte diese Entscheidung anders ausfallen.

6.2 Abhängigkeit der Metriken untereinander

Die Metriken für den Kontext finden früher im Ablauf einer Anfrage an das RAG statt. Sollten diese Metriken besonders gut oder schlecht sein, lässt sich deutlich sehen, wie sich dies auf die danach folgenden Metriken auswirkt.

Wenn die Metriken für den Kontext (context_precision und context_recall) eine Bewertung von 0 haben, ist die Wahrscheinlichkeit, dass die anderen Metriken auch 0 sind, relativ hoch. In diesem konkreten Beispiel werden die Abhängigkeiten des OpenAI RAGs für die 300 Fragen gezeigt. Es lässt sich sehen, dass die faithfulness deutlich weniger von den Kontext-Metriken abhängt.

In der finalen Auswertung sollten Anwender also überlegen, Metriken wie Answer Accuracy besonders zu betrachten, wenn die Kontext-Metriken gut bewertet worden sind.

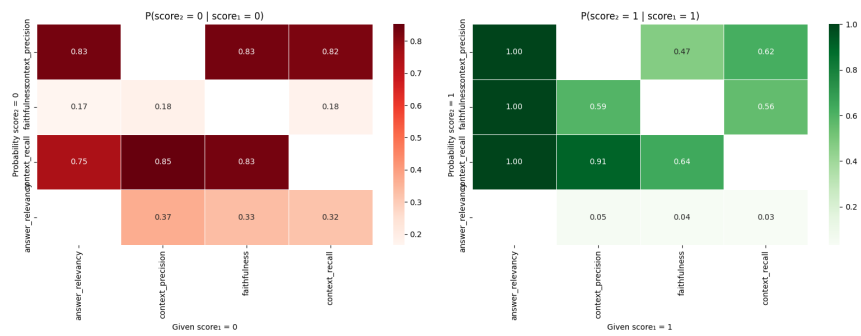


Abbildung 6.10: Abhängigkeit der Metriken voneinander (OpenAI, 300 Fragen, 400 Dokumente)

7 Zusammenfassungen

7.1 Benutzung von RAGAS

Dank der vielen Integrationen hat sich die Verwendung von RAGAS als einfach herausgestellt.

7.2 Fragebögen

Die Generierung von Fragebögen ist eines der Alleinstellungsmerkmale von RAGAS. Die Generierung der Fragebögen hat sich aus softwaretechnischer Sicht als unkompliziert erwiesen. Es gab zu den wichtigen Themen ausreichend Dokumentation und Beispiele.

Es ist mit RAGAS möglich, Fragebögen zu generieren, jedoch gibt es mehrere Faktoren, die Qualität und Praxistauglichkeit beeinflussen:

- Die Fähigkeit des LLMs, zuverlässig hochwertige Antworten zu generieren.
- Die Dokumente: Je komplexer und zusammenhangsloser die Dokumente sind, desto schlechter lassen sich Fragen generieren.

Bei der Generierung von Fragebögen mit DeepSeek kam es allein durch die nicht generierten oder zu irrelevanten Themen generierten Fragen zu einer Fehlerquote von bis zu 45 %. Selbst bei händisch ausgewählten Dokumenten lag die Fehlerquote bei mindestens 16 %.

Die händische Überprüfung hat dann weiter gezeigt, dass DeepSeek Probleme mit der konstanten Generierung von sinnvollen Fragen hat. Hier wiesen bis zu 65 % der Fragen für ungefilterte Dokumente Mängel auf! Selbst bei den gefilterten Dokumenten waren mindestens 30 % mangelbehaftet.

Die Generierung von Fragebögen mit OpenAIs GPT-4 hatte in Bezug auf nicht generierte oder Fragen zu irrelevanten Themen eine deutlich niedrigere Fehlerquote. Es gibt einen Ausreißer mit 20 %, der Rest bleibt jedoch deutlich unter 10 %. Die manuelle Auswertung hat hier aber auch gezeigt, dass viele Fragen, bis zu 60 % bei ungefilterten Dokumenten, Mängel aufweisen. Bei gefilterten Dokumenten kommt GPT-4 auf durchschnittlich 17,5 % und halbiert damit die Fehlerquote im Vergleich mit DeepSeek.

Die Qualität der Fragebögen ist entscheidend, da sich hier entstandene Fehler weiter bis in die Bewertung durchziehen und eine korrekte Bewertung des eigentlichen RAGs verzerren!

Die bei den Versuchen generierten Fragebögen lassen Zweifel an einer zuverlässigen und hochwertigen Generierung von Fragen aufkommen.

7.3 Bewertung

Auch das Generieren von Bewertungen hat sich mithilfe von RAGAS als einfach umsetzbar erwiesen. Sowohl das Tracing als auch die Kostenberechnung waren für die unterstützten Modelle problemlos zu benutzen. Das Tracing erlaubt außerdem einen tieferen Blick in die Berechnung der Metriken und macht das ganze System transparenter.

Es hat sich jedoch bei der manuellen Durchsicht gezeigt, dass hier bei DeepSeek 60 % und bei GPT-4 30 % der Fragen nicht richtig beantwortet wurden. Dies lässt sich teils auf die ungültigen Fragen in den Fragebögen zurückführen.

Bei den Metriken lässt sich sagen, dass die Metriken zum Kontext (`recall` und `precision`) gut abschneiden. Die `faithfulness` zeigt eine erhöhte Abweichung zur menschlichen Einschätzung und sollte mit einer Toleranz von 10 % beachtet werden.

Die `Answer Relevancy` hat die größte Abweichung; hier fällt auf, dass sowohl höhere als auch niedrigere Werte erwartet wurden.

7.4 Zuverlässigkeit

Sowohl die mehrfache gesamte Bewertung als auch das mehrfache Ausführen einzelner Metriken haben gezeigt, dass die Bewertung einer Schwankung von wenigen Prozentpunkten unterliegt. Es war zu sehen, dass LLMs mit mehr Parametern geringere Schwankungen aufwiesen. Hier war der Unterschied zwischen DeepSeek und OpenAI jedoch deutlich geringer als bei der Fragebogengenerierung oder der Bewertung. Insgesamt lässt sich sagen, dass die minimalen Schwankungen die Praxis-tauglichkeit nicht beeinflussen und sich auf die Ergebnisse verlassen werden kann.

7.5 Fazit

Insgesamt ist RAGAS kein kompletter Ersatz für die menschliche Bewertung von RAGs. Die Idee hinter RAGAS, Fragen ohne menschliches Zutun zu generieren, um Zeit zu sparen, ist mit besseren LLMs teilweise gelungen. Um jedoch ein aussagekräftiges und zuverlässiges

Ergebnis zu generieren, ist eine menschliche Kontrolle an mehreren Stellen notwendig. Zuerst bei der Auswahl der Dokumente: Hier muss sowohl ein Verständnis vorhanden sein, wie gut LLMs mit welchen Daten umgehen können, als auch welche Daten für das KMU relevant sind. Nach der Generierung der Fragebögen sollte erneut ein Mensch die Fragen überprüfen, um grob falsche Fragen zumindest zu löschen.

Da die Berichte relativ konstante Bewertungen abgeben, lassen sich dann durchaus Verschlechterungen oder Verbesserungen am RAG messen. Die Metriken geben Aufschluss darüber, welcher Teil des Systems nicht funktioniert; diese Zusammenhänge ließen sich sehr gut sehen.

Insgesamt muss jedoch auch der Zeit- und Kostenaufwand für eine solche Bewertung in Betracht gezogen werden. Für eine aktive Entwicklung ist das Abwarten von 17 Stunden für eine Bewertung eines RAGs nicht praxistauglich und ein Hindernis. Eine Bewertung innerhalb von einer Stunde ist praxistauglich, ist jedoch ein Kostenfaktor. Hier muss der Anwendungsfall genauer betrachtet werden.

7.6 Zukunftsausblick

Für Unternehmen bieten LLMs und RAGs großes Potenzial für Kosteneinsparungen; die Qualitätskontrolle spielt dabei eine immer größere Rolle. RAGAS bietet gute Ansätze, um die Qualitätskontrolle zu automatisieren. Dass RAGAS in Zukunft in die Prozesse zur Bewertung solcher Systeme einfließt, ist daher sehr wahrscheinlich.

7.7 Reflexion der Arbeit

Literatur

- [1] Beatrust. *RAG Evaluation: Assessing the Usefulness of Ragas*. Accessed: 2025. 2024. URL: https://tech.beatrust.com/entry/2024/05/02/RAG_Evaluation%3A_Assessing_the_Usefulness_of_Ragas.
- [2] Jiawei Chen u. a. „Benchmarking Large Language Models in Retrieval-Augmented Generation“. In: *ArXiv abs/2309.01431* (2023). URL: <https://arxiv.org/pdf/2309.01431>.
- [3] Try Chroma. *Try Chroma*. Accessed: 2025. 2025. URL: <https://www.trychroma.com/>.
- [4] DeepSeek-AI. „DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning“. In: *arXiv preprint arXiv:2501.12948* (2024).
- [5] Shahul Es u. a. „Ragas: Automated Evaluation of Retrieval Augmented Generation“. In: *arXiv preprint arXiv:2309.15217* (2023). DOI: 10.48550/arXiv.2309.15217. URL: <https://arxiv.org/abs/2309.15217>.
- [6] ESF. *Glossar*. Accessed: 2025. 2025. URL: https://www.esf.de/portal/DE/Service/Glossar/Functions/glossar.html?cms_lv3=f748ebe5-3f04-4af0-ae3f-64f888942114&cms_lv2=3943ee31-db5a-48c8-96e9-c69287930b3e.
- [7] ExplodingGradients. *Integrations – How-to guide*. Zugegriffen am 18. Juni 2025. Mai 2025. URL: <https://docs.ragas.io/en/latest/howtos/integrations/>.
- [8] Hugging Face. *Pleias-RAG-1B*. Accessed: 2025. 2025. URL: <https://huggingface.co/PleIAS/Pleias-RAG-1B>.
- [9] Luyu Gao u. a. „RT-RAG: Leveraging Retrieval-Generated Chains for Open-Domain Question Answering“. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023, S. 9784–9800. URL: <https://aclanthology.org/2023.acl-long.546/>.
- [10] Gemini Team. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf. Google DeepMind Technical Report. 2024.

-
- [11] Gemini Team Google. „Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context“. In: *arXiv preprint arXiv:2403.05530* (2024).
 - [12] Gemini Team Google. „Gemini: A Family of Highly Capable Multimodal Models“. In: *arXiv preprint arXiv:2312.11805* (2023).
 - [13] Harrison Chase and LangChain contributors. *LangChain: Language model application development framework*. Erste Veröffentlichung Oktober 2022; Version aktuell am 26. Juni 2025. 2022. URL: <https://www.langchain.com/>.
 - [14] Thorsten Honroth, Julien Siebert und Patricia Kelbert. *Retrieval Augmented Generation (RAG): Chatten mit den eigenen Daten*. Zugriff am 7. Februar 2025. Mai 2024. URL: <https://www.iese.fraunhofer.de/blog/retrieval-augmented-generation-rag/>.
 - [15] Nomic Team. *Introducing Nomic Embed: A Truly Open Embedding Model*. <https://www.nomic.ai/blog/posts/nomic-embed-text-v1>. Online; veröffentlicht auf dem Nomic AI Blog. Feb. 2024.
 - [16] Ollama. *Ollama*. Accessed: 2025. 2025. URL: <https://ollama.com/>.
 - [17] OpenAI. „GPT-4 Technical Report“. In: *arXiv preprint arXiv:2303.08774* (2023). URL: <https://arxiv.org/abs/2303.08774>.
 - [18] Luka Panic. *RAG in der Praxis – Generierung synthetischer Testdatensätze*. Abgerufen am 30. Mai 2025. 2024. URL: <https://pixon.co/blog/rag-in-practice-test-set-generation>.
 - [19] Ofir Press u. a. „Measuring Faithfulness in Chain-of-Thought Reasoning“. In: *arXiv preprint arXiv:2211.08411* (2022). URL: <https://arxiv.org/abs/2211.08411>.
 - [20] Ragas. *Context Entities Recall*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_entities_recall/.
 - [21] Ragas. *Context Precision*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_precision/.
 - [22] Ragas. *Context Recall*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall/.
 - [23] Ragas. *Faithfulness*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness/.
 - [24] Ragas. *Metrics*. Accessed: 2025. 2025. URL: <https://docs.ragas.io/en/stable/concepts/metrics/>.

-
- [25] Ragas. *Noise Sensitivity*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/noise_sensitivity/.
 - [26] Ragas. *Nvidia Metrics*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/nvidia_metrics/.
 - [27] Ragas. *Query types in RAG*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/test_data_generation/rag/#query-types-in-rag.
 - [28] Ragas. *Response Relevancy*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/answer_relevance/.
 - [29] Ammar Shaikh u. a. *CBEval: A framework for evaluating and interpreting cognitive biases in LLMs*. 2024. DOI: 10.48550/arXiv.2412.03605. arXiv: 2412.03605 [cs.CL]. URL: <https://arxiv.org/abs/2412.03605>.
 - [30] Doit Software. *ChatGPT Statistiken*. Accessed: 2025. 2025. URL: <https://doit.software/de/blog/chatgpt-statistiken#screen5>.
 - [31] Hugo Touvron u. a. „Llama 2: Open Foundation and Fine-Tuned Chat Models“. In: *arXiv preprint arXiv:2307.09288* (2023).
 - [32] Ashish Vaswani u. a. „Attention Is All You Need“. In: *CoRR* abs/1706.03762 (2017). Preprint. URL: <https://arxiv.org/abs/1706.03762>.
 - [33] Wikipedia. *Confusion matrix*. Accessed: 2024. 2024. URL: https://en.wikipedia.org/wiki/Confusion_matrix.
 - [34] Jingfeng Yang u. a. *Large Language Models are not Fair Evaluators*. 2023. arXiv: 2305.17926 [cs.CL]. URL: <https://arxiv.org/abs/2305.17926>.

Anhang

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Anmerkung: In einigen Studiengängen findet sich die Erklärung unmittelbar hinter dem Deckblatt der Arbeit.

Ort, Datum

Unterschrift