
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*

Eine Kurzfassung (wenn verlangt) in Deutsch und/oder in Englisch (*Abstract*) umfasst auf etwa 1/2 bis 1 Seite die Darstellung der Problemstellung, der angewandten Methode(n) und des wichtigsten Ergebnisses.

Wie man ein gelungenes Abstract verfasst, erfahren Sie in den Seminaren oder der Beratung des Schreibzentrums der Kompetenzwerkstatt¹.

Schlagwörter/Schlüsselwörter: gegebenenfalls Angabe von 3 bis 10 Schlagwörtern.

¹<https://www.th-koeln.de/schreibzentrum>

Inhaltsverzeichnis

Tabellenverzeichnis	IV
Abbildungsverzeichnis	V
1 Einleitung	1
1.1 Wie funktioniert ein RAG	2
1.1.1 Vorteile von RAGs	2
1.1.2 Kompetenz des Betreibers	3
1.1.3 Datenbasis	3
1.1.4 Budget	4
1.2 Objektive Beurteilung von RAGs	4
1.3 Darstellung des Themas und der Forschungsfragen	4
1.4 Praxistauglichkeit und Herausforderungen	4
1.5 Softwaretechnische Fragestellungen	5
1.6 Rechtliche Fragestellungen	6
2 Methoden und Materialien	7
2.0.1 Werkzeuge	7
2.0.2 Daten	8
2.0.3 Fragebögen	9
2.0.4 Evaluation	10
2.1 Metriken	11
2.1.1 Retrieval Augmented Generation	11
2.1.2 Nvidia Metrics	13
2.1.3 Natural Language Comparison	14
2.1.4 Non LLM String Similarity	15
2.1.5 General purpose	16
2.1.6 Andere Metriken	16
2.1.7 Irrelevante Metriken	16
3 Ähnliche Arbeiten	17
3.1 RAG Evaluation: Assessing the Usefulness of Ragas	17
3.2 RAG-Bewertungsprozess	18

4	Versuche	20
4.1	Versuchsplan	20
4.1.1	Forschungsfragen	20
4.1.2	Variablen in den Versuchen	20
4.1.3	Kosten- und Zeitanalyse	22
4.1.4	Versuchsprotokoll	22
4.1.5	Bewertungskriterien für die Geschäftstauglichkeit	22
4.2	Konkretisierung der versuche	24
4.2.1	Dokumentenverarbeitung	24
4.2.2	Testset-Generierung	24
4.2.3	Bewertung	24
5	Ergebnisse und Diskussionen	26
5.1	Ergebnisse aus den Versuchen	26
5.1.1	Generierte Fragebögen	26
5.1.2	Manuelle Auswertung der Fragebögen	27
5.1.3	Auswertung der Reports	29
5.1.4	Unterschiede über mehrere Durchläufe	32
5.1.5	Zuverlässigkeit von Metriken	34
5.1.6	Kostenberechnung	37
5.2	Abhängigkeit der Metriken untereinander	38
5.3	Identifikation von Interessenten	38
6	Zusammenfassungen	39
6.1	Benutzung von RAGS	39
6.2	Testsets	39
6.3	Bewertung	40
6.4	Fazit	40
6.5	Zukunftsausblick	41
6.6	Reflektieren der Arbeit	41
	Literatur	42
	Anhang	44

Tabellenverzeichnis

4.1	Kombinationen aus Dokumentanzahl und Embedding-Modell für die Versuche (X = Kombination wird getestet)	24
4.2	Kombinationen aus Dokumentanzahl und Testset-Größe	24
4.3	Übersicht aller 24 zu generierenden Bewertungsberichte mit Abkürzungen und Wiederholungen	25
5.1	Übersicht der generierten Fragen und Fehlerraten pro Testset für DeepSeek	27
5.2	Übersicht der generierten Fragen und Fehlerraten pro Testset	27
5.3	Anzahl fraglicher Fragen pro Testset und Gesamtübersicht für Ollama . . .	28
5.4	Anzahl fraglicher Fragen pro Testset und Gesamtübersicht für OpenAI . .	29
5.5	Verteilung der Bewertungen für DeepSeek (mit Prozentangaben)	30
5.6	Verteilung der Bewertungen für OpenAI (gesamt) mit Prozentangaben . .	31
5.7	Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für DeepSeek	34
5.8	Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für GPT-4	34
5.9	Dauer der Evaluation pro Dokumentenzahl mit OpenAI	36
5.10	Dauer der Evaluation pro Dokumentenzahl	37

Abbildungsverzeichnis

1.1	Struktur eines RAGs, Quelle: [4]	2
3.1	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.	19
5.1	DeepSeek Ergebnis für 300 Fragen	31
5.2	ChatGPT Ergebnis für 300 Fragen	31
5.3	Abweichungen des Faithfulness Scores.	32
5.4	DeepSeek Ergebnis für 100 Fragen	33
5.5	ChatGPT Ergebnis für 100 Fragen	33
5.6	Bewertung der vier Durchläufe mit DeepSeek	33
5.7	Bewertung der vier Durchläufe mit GPT-4	34
5.8	Abweichungen des Answer Relevancy Scores.	35
5.9	Abweichungen des Faithfulness Scores.	36
5.10	Abweichungen des Answer Relevancy Scores.	38

1 Einleitung

Im Jahr 2022 veränderte OpenAI mit ihrem browser basierten ChatGPT (Generative Pre-trained Transformer) die Welt komplett. In nur fünf Tagen erreichte ChatGPT die eine Millionen Nutzer und ist im Leben vieler Menschen Alltag und in manchen gar nicht mehr weg zu denken. Die GPT KI (Künstliche Intelligenz) von OpenAI gehört zu der Familie der LLMs (Large Language Models) oder auch MLLMs (Multimodal Large Language Models), wenn diese weitere Datenmodalitäten verarbeiten können wie zum Beispiel Bilder, Audi und Video. Inzwischen haben LLMs von anderen Anbietern mit der Qualität und den Fähigkeiten von OpenAIs GPT gleichgezogen. Inzwischen 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 Google's Gemini sind nur über eine API (Programmierschnittstellen) meistens gegen Entgeltung verfügbar. Open-Source Modelle wie Liang Wenfengs DeepSeek oder Metas LLAMA erfreuen sich immer größerer Beliebtheit da sie gratis auf der eigenen Hardware ausgeführt werden können.

Im Oktober 2023 kam ich das erste mal mit RAGs (Retrieval-Augmented Generation) in Kontakt, damals war die Idee über mehrerer Firmeninterne Informationsquellen mithilfe eines LLMs Fragen zu beantworten. Bei einem Hackathon gelang es uns einen Prototypen zu entwickeln der mit einem gewissen Erfolg Fragen zu Firmeninternen Themen beantworten konnte.

Einer der Schritte während der Entwicklung war das ständige Testen der neusten Änderungen um zu gucken, ob das System noch funktioniert und oder wie es sich verschlechtert hat. Das war schon damals immer relativ mühselig und raubte uns wertvolle Zeit diese Bewertungen vor zu nehmen. Gerne hätten wir unterschiedliche Prompts innerhalb unseres Systems getestet oder automatisch eine Überprüfung unseren neusten Änderungen gehabt.

RAGAs wurde entwickelt um diese Probleme zu lösen, es hat zudem das Alleinstellungsmerkmal, dass man weder eigenen Fragen noch die generierten Fragen selber beantworten muss. Sowohl die Generierung eines Fragenkataloges (Testsets) als auch die Beantwortung der Fragen um eine Musterlösung zu haben nimmt RAGAS mit Hilfe von LLMs vor. Mithilfe dieses Testsets und von RAGAS eigens entwickelten Metriken welche die wichtigsten Funktionen eines RAGs abdecken kann dann eine Bewertung des Systems vorgenommen werden.

Damit benutzt RAGAS die neue LLM Technologie um das durch LLMs entstandene System selber zu testet. Dies spart viele Menschliche Ressourcen welche Zeit und Kostenintensiv sind.

1.1 Wie funktioniert ein RAG

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

[4]

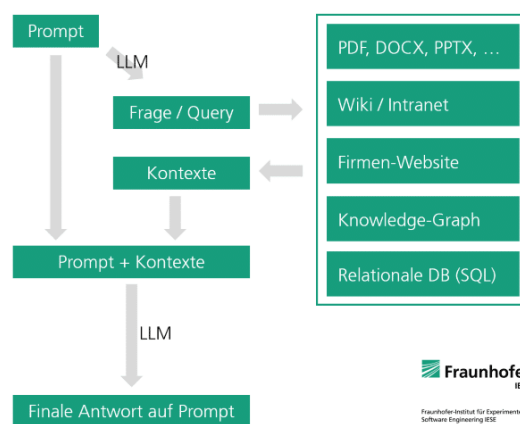


Abbildung 1.1: Struktur eines RAGs, Quelle: [4]

1.1.1 Vorteile von RAGs

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

1. Im Trainingsset für die LLMs selten vorkommendes Wissen können selbst LLMs schlecht lernen. [2] [6]
2. LLMs haben einen gewissen Wissensstand und müssen weiter trainiert werden, um die neuesten Informationen zu kennen.

3. Firmen interne Dokumente sind nicht im Trainingsset und daher können LLMs keine Fragen zu Firmen Internen Daten beantworten.

Die Nutzung eines RAGs ist eine der drei Möglichkeiten, um ein LLM zu verbessern. RAGs haben neben dem Fine-Tuning und dem nutzen eines LLMs mit großem Sichtfenster entscheidende Vorteile.

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

1.1.2 Kompetenz des Betreibers

Für das Finetuning von LLMs ist ein gewisses 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 Verständniss von LLMs und effektiven Methoden für den suchenden (Retrival) Teil des RAGs notwendig. Zudem müssen hier manuell für jeden Datentypen (Email, PDF etc.) angebunden werden. Sollte ein seltender Datentyp verwendet werden muss hier eventuell eigens eine Anbindung entwickelt werden.

1.1.3 Datenbasis

Sollten die Daten dynamisch sein, ist das RAG die vorzuziehende Lösung. Durch die Eigenschaften der schnellen und kontinuierlich aktualisierung der Daten. Wie vorhin erläuteter 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.

1.1.4 Budget

Das Finetuning erfordert für lange Zeit teure Rechenzeit auf Hochleistung-GPUs, was das Training eines Modells kostenintensiv macht.

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

Die wohl kostenintensivste Methode ist das Nutzen eines LLMs mit einer großen Context Window.

1.2 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, diesen Prozess unter anderem mithilfe von LLMs zu automatisieren. Diese Tools generieren aus den ihnen gegebenen Daten Fragebögen, die auf eine 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 zum Beispiel 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.

1.3 Darstellung des Themas und der Forschungsfragen

In dieser Bachelorarbeit wird untersucht, wie gut diese Tools sowohl subjektive als auch objektive Bewertungen durchführen können. Im Mittelpunkt werden die beiden Tools RAGAS und Giskard stehen, welche die Bewertung durchführen.

1.4 Praxistauglichkeit und Herausforderungen

Es stellen sich mehrere Herausforderungen für die Bewertung von RAGs durch diese Tools.

- Die Kosten, die bei der Bewertung entstehen.

- Die Zeit, welche es dauert, die Bewertung durchzuführen, die Bewertung kann schneller durchgeführt werden kann, 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 auch auf dem neuesten Stand gehalten werden, da sich dieses noch relativ junge Thema schnell entwickelt.

1.5 Softwaretechnische Fragestellungen

In dem Artikel *RAG in der Praxis – Generierung synthetischer Testdatensätze* untersucht Luka Panic [5] 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 diesem Paper [17] wird gezeigt, dass, wenn ein LLM eine von zwei gegebenen Antworten aussuchen müsste, die erste bevorzugt wurde, selbst wenn die gleiche Frage mit anderer Reihenfolge gestellt wurde. RAGs vergleicht keine Antworten miteinander und daher ist dieser Bias kein direktes Problem für uns. Was jedoch einen Einfluss auf die Bewertung von Antworten haben kann, ist der Bias zu gewissen Nummern. Wie in [15] beschrieben, bevorzugen LLMs bei der Bewertung lieber Zahlen, welche Mehrfache von 5 und 10 sind.

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

Wie in diesem Paper [3] 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 sind 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.

<https://huggingface.co/PleIAs/Pleias-RAG-1B>

Pitfalls in LLM Assisted Evaluation <https://medium.aiplanet.com/evaluate-rag-pipeline-using-ragas-fbdd8dd466c1>

1.6 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 Nutzer*innen der RAGs müssen sich den aus der KI-VO ergebenden Pflichten bewusst sein, wie bei der Verwendung von vertraulichen Daten.

2 Methoden und Materialien

2.0.1 Werkzeuge

Für die RAGs und die Bewertung der RAGs werden Tools benötigt, im nachfolgenden werden diese Tools genauer erklärt.

Ollama

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.

Vektordatenbank

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.

RAGAS

RAGAS ist eine Bibliothek, die Werkzeuge bereitstellt, um die Evaluation von Large Language Model (LLM) Anwendungen zu verbessern. Sie wurde entwickelt, um die Bewertung von LLM-Anwendungen einfach und zuverlässig zu gestalten.¹

RAGAS ist ein Open-Source-Tool und liefert neben dem Tool selber hilfreiche Dokumentation für die Metriken und die Bewertung von RAGs. Es werden Funktionen wie die automatische Generation 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.

¹<https://docs.ragas.io/en/stable/#frequently-asked-questions>

Was RAGAS von den vorherigen Tools unterscheidet, ist, dass keine "reference answer" benötigt wird. RAGAS ist beliebt, da es sich gut mit vielen Tools integriert. [1]

Giskard

Giskard ist ein teils Open-Source-Tool, welches die Bewertung von RAGs unterstützt. Der Schwerpunkt von Giskard liegt größtenteils auf der generellen Bewertung von LLMs. Dazu gehören unter anderem Prompt-Injectionen, Halluzinationen und andere Fehler, die durch die Verwendung von LLMs entstehen können.

2.0.2 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 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 eigenen Unternehmungen ließen sich 73 Dokumente finden die nutzbar sind. 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 auch 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.

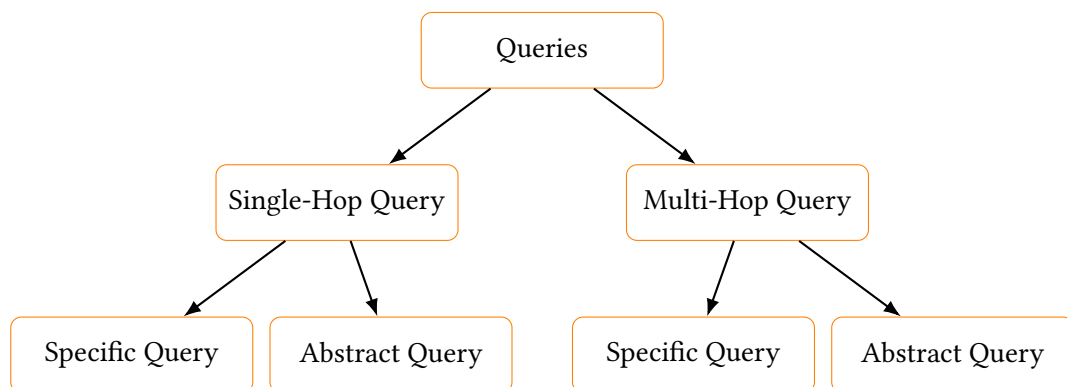
Die folgenden drei Stufen wurden gewählt, um typische Anwendungsszenarien realistisch abzubilden:

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.

2.0.3 Fragebögen

Frageotypen

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



Diese verschiedenen Frageotypen 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 [13] wird für konkrete Fragen das Beispiel gemacht "Wann hat Einstein die Relativitätstheorie veröffentlicht?" während eine abstraktere Frage wäre, "Wie hat Einsteins Relativitätstheorie unser Verständnis der Welt verändert?"

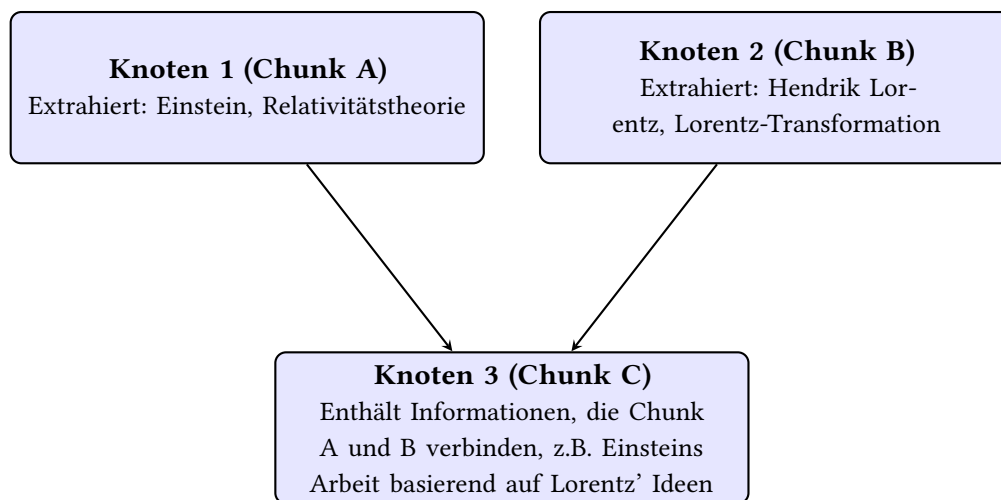
Bei Multihop Querys handelt es sich um Fragen, welche mehr als eine Wissensabfrage benötigen, die Frage "Welche Wissenschaftler haben Einsteins Relativitätstheorie beeinflusst und welche Theorie haben sie vorgeschlagen?" benötigt erst eine Abfrage um die Wissenschaftler herauszufinden und dann weitere um die jeweils vorgeschlagene Theorie abzufragen. Für die Abstrakte Multihop 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.

Wissensgraf

Für die eben erwähnten Multihop Querys müssen aus den gegebenen Dokumenten Themen, welche zusammenhängen aber nicht direkt im gleichen Dokument sind, gefunden werden. Da dies bei großen Datensätzen manuell oder selbst mit einem LLM schwierig ist, wird ein Wissensgraf erstellt.

Dies passiert in drei Schritten, zuerst werden die Dokumente beim sogenannten chunking in kleiner 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. (Vergleich mit Wikipedia Links in Artikeln)



Für die Daten aus den Versuchen mit 100 Dokumenten hat RAGAS 27 Themen identifiziert, unter anderem waren dort folgende Themen dabei:

Finanzmanagement, Bildungsprojekt Digitalisierung, Projektmanagement und Planung, Zuwendungsverwaltung, Break-Even-Analyse, Finanzplanung und Investitionen, Finanzplanung und Liquidität

2.0.4 Evaluation

2.1 Metriken

In diesem Kapitel geht es um die verschiedenen Metriken, die für die Bewertung von RAG Evaluationstools verwendet werden können. Metriken sind das Herzstück der Bewertung von RAGs, da sie die Qualität des RAGs bewerten und somit die Entwicklung und den Fortschritt des RAGs messen.

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

2.1.1 Retrieval Augmented Generation

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

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. Die Präzision@k ist das Verhältnis der Anzahl relevanter Textabschnitte auf Rang k zur Gesamtanzahl der Textabschnitte auf Rang k. (eigene Übersetzung nach [8])

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

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 [9])

Wenn es eine gute 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.

Context Entities Recall

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

Die ContextEntityRecall-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 [7])

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

Noise Sensitivity

NoiseSensitivity 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 [11])

Diese Metrik ist eine der wichtigsten Metriken in dieser Arbeit da sie die Richtigkeit der Antworten und damit die Qualität des RAGs bewertet.

Response Relevancy

Die ResponseRelevancy-Metrik misst, wie relevant eine Antwort im 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 [14])

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 zu der Frage sind.

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:

$$\text{Faithfulness Score} = \frac{\text{Anzahl der durch den Kontext gestützten Aussagen in der Antwort}}{\text{Gesamtanzahl der Aussagen in der Antwort}} \quad (2.1)$$

(eigene Übersetzung nach [10])

Multimodal Faithfulness/Multimodal Relevance

Da sich diese Metriken mit mehr als textuellen Daten befassen, werden diese nicht im Rahmen dieser Arbeit betrachtet.

2.1.2 Nvidia Metrics

Diese Metriken sind subjektiver Art und benutzen wieder eine LLM, um die Bewertung zu treffen. Hier werden einzelne Bewertungen generiert, welche keinen tieferen Einblick in die Bewertung gewähren.

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 "LLM-as-a-judgePrompts, die jeweils eine Bewertung (0, 2 oder 4) zurückgeben. 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 [12])

Das LLM bewertet die Antwort mit der Referenz und auch die Referenz mit der Antwort. Hat Vorteile gegenüber der Answer Correctness, da es weniger Aufrufe mit weniger Tokens an LLM braucht. Es werden im Vergleich zur Answer Correctness auch robustere Bewertungen getroffen, bietet jedoch weniger Einblicke in die Bewertung. Diese Metrik wird im Rahmen dieser Arbeit betrachtet auch um einen Vergleich zu anderen Metriken zu haben.

Context Relevance

Diese Metrik ist sehr ähnlich zur Context Precision, als Alternative und um einen Vergleich zu haben wird diese im Rahmen dieser Arbeit betrachtet, auch wenn sie keine direkte Aussage über das zu bewertende LLM macht.

Response Groundedness

Wenn die Answer Accuracy eine gute Bewertung liefert, ist die Response Groundedness eine gute Bewertung für die Faktualität der Antwort. Diese Logik ist ähnlich zur Kombination von Context Relevancy und Context Precision. Hier wird es in den Versuchen interessant zu vergleichen wie diese Metriken zusammenhängen.

2.1.3 Natural Language Comparison

Factual Correctness

Diese Metriken basieren zu Teilen auf der Wahrheitsmatrix (Confusion matrix), welche die vier Kategorien True Positive, False Positive, False Negative und True Negative definiert.[16] Aus dieser Matrix lassen sich dann precision, recall und f1 score berechnen.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2.2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (2.3)$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.4)$$

[16]

Semantic Similarity

This metric uses embeddings to calculate the semantic similarity between the answer and the reference. TODO: should this be used?

2.1.4 Non LLM String Similarity

Wie der Name schon sagt, wird die String Similarity ohne LLM berechnet. Diese Metriken sind relative einfache Metriken und werden im Rahmen dieser Arbeit keine große Rolle spielen, jedoch als Vergleich zu anderen Metriken dienen.

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.

ROUGE Score

Mithilfe von n-gram recall, precision, und dem F1 score wird die Ähnlichkeit zwischen der Antwort und der Referenz berechnet.

String Presence

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

Exact Match

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

2.1.5 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
- Simple Criteria Scoring
- Rubrics based Scoring
- Instance Specific Rubrics Scoring

2.1.6 Andere Metriken

Summarization

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

2.1.7 Irrelevante Metriken

SQL

SQL spezifische Metriken, welche nicht im Rahmen dieser Arbeit betrachtet werden.

Agents or Tool use cases

Metriken zum Bewerten des Einsatzes von Agenten oder Tools, dies liegt ebenso außerhalb des Themas dieser Arbeit. <https://docs.ragas.io/en/stable/concepts/metrics/>

Diese Metrik wird Teil dieser Arbeit sein, da sie in gewissen Nutzungsfällen, wie z.B. stark Fakten basierte Fragen, eine gute Bewertung liefern kann.

3 Ähnliche Arbeiten

3.1 RAG Evaluation: Assessing the Usefulness of Ragas

https://tech.beatrust.com/entry/2024/05/02/RAG_Evaluation%3A_Assessing_the_Usefulness_of_Ragas

Das Team von Beatrust hat im Februar 2024 eine Reihe zu RAGs veröffentlicht. Es werden unter anderem die Notwendigkeit und auch die einzelnen Metriken von RAGAS erklärt. Im dritten Artikel dieser Reihe machen sie ein Versuch um die Nützlichkeit von RAGAS zu untersuchen. Der Versuch besteht aus 50 Fragen aus einem Interessensfeld des Authors, diese wurden vom einem RAG mit GPT-4 und einem mit GPT-3.5-turbo beantwortet und dann sowohl von RAGAS als auch von ihm bewertet. Der Author kommt zu dem Ergebniss, dass RAGAS geeignet ist um RAGs zu bewerten und besser ist als die Bewertung von Langchain. Es wird jedoch angemerkt, dass der Author eine höhere Übereinstimmung mit seinen Ergebnissen erwartet hätte.

<https://www.qed42.com/insights/simplifying-rag-evaluation-with-ragas>

<https://medium.aiplanet.com/evaluate-rag-pipeline-using-ragas-fbdd8dd466c1>

<https://arxiv.org/pdf/2309.01431>

3.2 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 des Testsets

- 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.

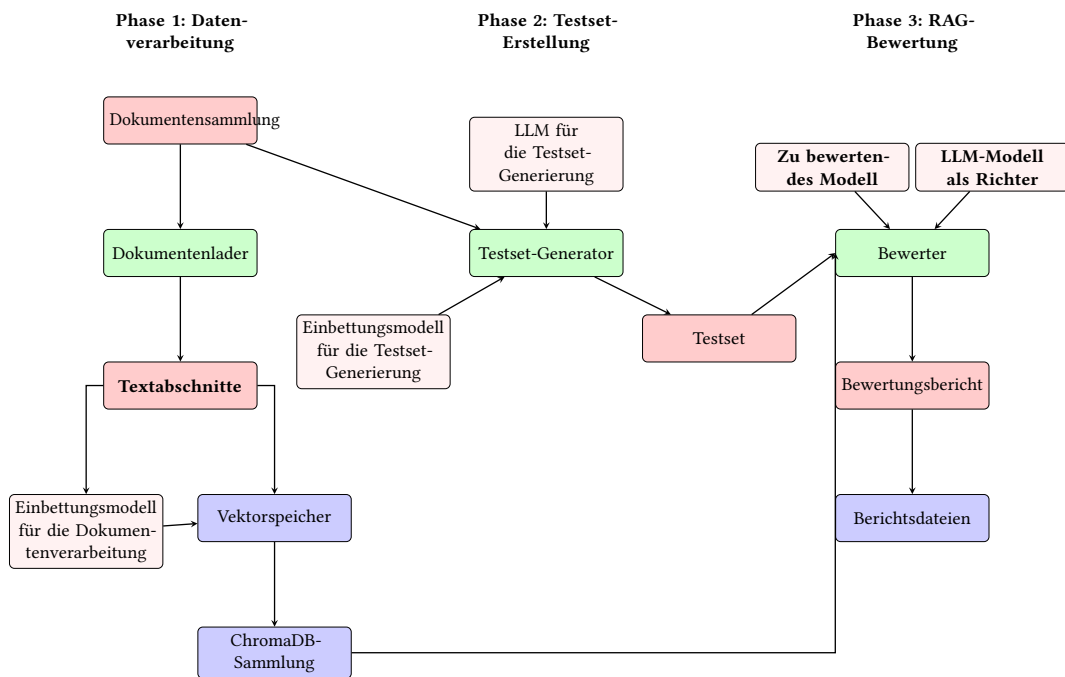


Abbildung 3.1: Flussdiagramm des RAG-Bewertungsprozesses, das die Interaktion zwischen verschiedenen Komponenten und Modellen zeigt. Spezifische Modellnamen (z.B. gpt-4-turbo, text-embedding-3-large) sind im Haupttext beschrieben.

4 Versuche

4.1 Versuchsplan

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

4.1.1 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 kleinen Unternehmen geeignet?
2. Wie beeinflussen verschiedene Dokumenttypen 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 kleinen Unternehmen?

4.1.2 Variablen in den Versuchen

Dokumenttypen

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

- PDF (.pdf)
- Klartext (.txt)
- 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** 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 kleine Unternehmen 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, was rechtliche Vorteile bietet, aber auch Nachteile. Entscheidend ist zudem wieder die Technische Kompetenzen welche benötigt wird um diese Modelle selber 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. Sie haben ähnliche Vor- und Nachteile wie die Mittelklasse-API-Modelle.

Bewertungsmetriken

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

4.1.3 Kosten- und Zeitanalyse

Ob wir dies tun wollen, ist noch nicht klar. RAGAS bietet Kostenberechnung an, aber ich habe es mir noch nicht angesehen.

4.1.4 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 und zunächst 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.

4.1.5 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 kleinen Unternehmen sind, mit klaren Anleitungen zu optimalen Konfigurationen und Implementierungsstrategien.

4.2 Konkretisierung der versuche

4.2.1 Dokumentenverarbeitung

Damit die Dokumente in der Vektordatenbank gesichert werden können müssen sie erst zu Vektoren konvertiert werden. Das wird mit Embeddings gemacht, hier werden wir auch wieder die von OpenAI verwenden aber auch eine Open-Source Variante von nomic.ai.

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

Tabelle 4.1: Kombinationen aus Dokumentanzahl und Embedding-Modell für die Versuche
(X = Kombination wird getestet)

4.2.2 Testset-Generierung

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

Dokumentanzahl	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 4.2: Kombinationen aus Dokumentanzahl und Testset-Größe

4.2.3 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	1/1
2	OAI-E	10	30	GPT-4	GPT-4	4/4
3	OAI-E	100	50	GPT-4	GPT-4	1/1
4	OAI-E	100	100	GPT-4	GPT-4	4/4
5	OAI-E	400	150	GPT-4	GPT-4	1/1
6	OAI-E	400	300	GPT-4	GPT-4	1
7	OAI-E	10	15	DSK-R	DSK-R	1
8	OAI-E	10	30	DSK-R	DSK-R	1
9	OAI-E	100	50	DSK-R	DSK-R	1
10	OAI-E	100	100	DSK-R	DSK-R	1
11	OAI-E	400	150	DSK-R	DSK-R	1
12	OAI-E	400	300	DSK-R	DSK-R	1

Tabelle 4.3: 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

5 Ergebnisse und Diskussionen

5.1 Ergebnisse aus den Versuchen

5.1.1 Generierte Fragebögen

Da insgesamt 1.290 Fragen generiert werden lassen sich diese aufgrund des zeitlichen Aufwandes nicht alle bewerten. Aus jedem Fragebogen 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 (11 %) überhaupt generiert, der Rest ist aufgrund von technischen Problemen oder ungültigen Antworten seitens DeepSeek nicht generiert worden
- Von diesen sind 101 zu Themen rund um Bezahlmethoden, Versand und ähnliches. In den Dokumenten welche dem DeepSeek zu verfügung gestellt wurden traten diese Themen nicht auf. Diese Fragen sind daher als ungültig bewertet worden.

Am Ende bleiben also 166 von 300 Fragen übrig, **ca. 45 %** der angeforderten Fragen sind nicht daher irrelevant!

Beim generieren des Testsets mit 100 Fragen sah das ganze etwas besser aus

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

Am Ende haben wir für das testset mit 100Fragen eine Fehlerquote von **16 %**.

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

Angefragt	Generiert	Irrelevant	Verlust	Fehlerrate
15	11	0	4	27 %
30	27	7	3	33 %
50	40	0	10	20 %
100	88	4	12	16 %
150	137	49	13	41 %
300	267	101	33	45 %

Tabelle 5.1: Übersicht der generierten Fragen und Fehlerraten pro Testset für DeepSeek

Open-AI

Angefragt	Generiert	Irrelevant	Verlust	Fehlerrate
15	12	0	3	20%
30	30	1	0	3%
50	48	0	2	4%
100	95	1	5	6%
150	150	8	0	5%
300	300	16	0	5%

Tabelle 5.2: Übersicht der generierten Fragen und Fehlerraten pro Testset

5.1.2 Manuelle Auswertung der Fragebögen

DeepSeek

Bei der manuellen Sichtung der Testsets wurden weitere Fehler entdeckt

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.

Bei dieser Frage hat das LLM verdreht, wer bezahlt und fragt, wie viel die Schulen weniger **verdienen** und nicht wie viel sie weniger **bezahlen**.

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.

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 oder auch wie viel Prozent Erreichbarkeit garantiert wird.

Ebenso ist "Wie hoch ist die Gesamtsumme der Passiva?" eine Frage welche nicht spezifiziert um welches Jahr es sich handelt Fehleranfällig.

Es gab auch Fragen, welche sich vom Kontext verwirren lassen haben.

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.

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

Tabelle 5.3: Anzahl fraglicher Fragen pro Testset und Gesamtübersicht für Ollama

Die Fehlerquote von 45% zeigt, dass die fragen die DeepSeek generiert nicht einfach eingesetzt werden können. Es sind hier eindeutige Unterschiede im Vergleich zu von Menschen generierten Fragen erkenntlich.

OpenAI

Bei ChatGPT wurden auch Mängel bei der manuellen Überprüfung festgestellt.

Die Fragen werden so gestellt, dass sie den "gegebenen Kontext" bewerten sollen. Es fehlen

dadurch wichtige Informationen welche zum Finden der relevanten Dokumente fehlen.
 “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.”

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.

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	Fragliche Fragen
OpenAi – 10 Dok (10)	0
OpenAi – 10 Dok (30)	3
Summe 10 Dok: 3 / 20 = 15%	
OpenAi – 100 Dok (50)	3
OpenAi – 100 Dok (100)	1
Summe 100 Dok: 4 / 20 = 20%	
OpenAi – 400 Dok (150)	5
OpenAi – 400 Dok (300)	7
Summe 400 Dok: 12 / 20 = 60%	
Gesamt (OpenAi): 19 / 60 = 31,67%	

Tabelle 5.4: Anzahl fraglicher 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!

Bei einem Vergleich mit einem von Menschen erstellten Fragebogen sind hier jedoch deutlichen Unterschiede was die Qualität der Fragen betrifft.

5.1.3 Auswertung der Reports

Für die Auswertung der Reports werden wieder die selben Fragen wie vorher aus den Testsets verwendet. Dabei wird geguckt

- 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 Answere 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 % erst garnicht als sinnvoll sind.

Auch die nicht bewerteten Metriken sind mit bis zu 83 % fast 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	34 (56.7%)	26 (43.3%)	–
Gültige Antwort	22 (36.7%)	38 (63.3%)	–
context_precision	10 (16.7%)	–	50 (83.3%)
faithfulness	11 (18.6%)	2 (3.4%)	47 (78.0%)
context_recall	60 (100%)	–	–
answer_relevancy	43 (71.7%)	15 (25.0%)	2 (3.3%)
answer_relevancy sollte höher sein	4 (100%)	–	–

Tabelle 5.5: 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 Werte führen würden. Es kam jedoch zu zwischenzeitlichen RateLimits, diese könnten von einer Firma jedoch bei einem Vertragsschluss mit OpenAI erhöht werden.

GPT-4.1 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.

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 5.6: Verteilung der Bewertungen für OpenAI (gesamt) mit Prozentangaben

Probleme mit Code

Da es sowohl bei der Testset generierung als auch bei der Bewertung der Testsets, die 400 Dokumente nutzten, höhere Fehlerraten zu beobachten sind wird das genauer untersucht.

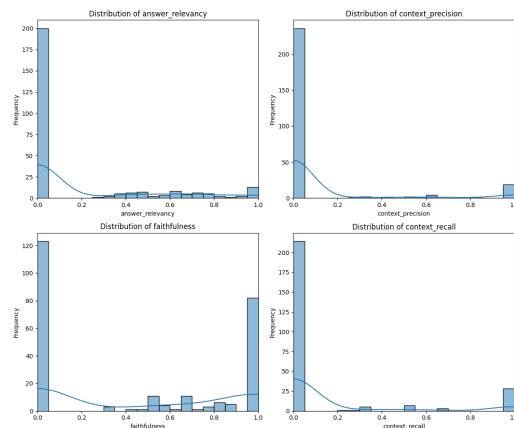


Abbildung 5.1: DeepSeek Ergebnis für 300 Fragen

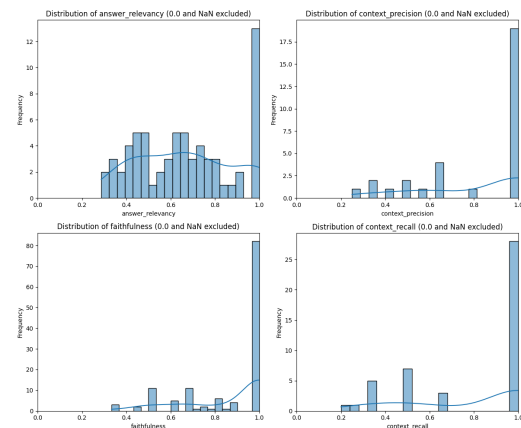


Abbildung 5.2: ChatGPT Ergebnis für 300 Fragen

Der Vergleich der Anzahl der 0.0 Bewertungen mit DeepSeek (Abbildung 5.1) im Vergleich zu OpenAI (Abbildung 5.2) ist eindeutig.

106 (40 %) der Ergebnisse insgesamt 276 zu bewertenden Fragen waren mit komplett 0.0 bewertet worden. Bei der Analyse der Speziellen Charaktere im Kontextes fällt auf, dass 89 Bewertungen (83 %) zu mehr als 5 % nur aus diesen besteht, 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 durch Dokumente mit minderer Qualität, im Bezug

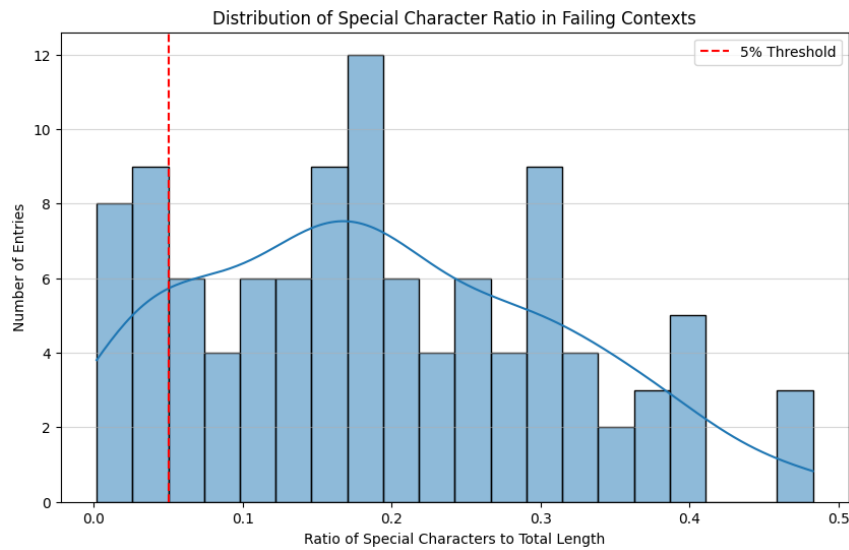


Abbildung 5.3: Abweichungen des Faithfulness Scores.

auf mögliche Fragestellungen, verwirren lassen. es zeigt sich wieder einmal, dass die Qualität der Daten eine entscheidende Rolle spielt! Wenn wir uns jetzt die Ergebnisse ohne Dokumente welche Code enthalten angucken sehen wir, dass die 0.0 Bewertungen bei DeepSeek deutlich zurück gehen aber wie zu erwarten OpenAi's ChatGPT-4 immer noch deutlich besser ist.

5.1.4 Unterschiede über mehrere Durchläufe

Um zu gucken, 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, 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 sind und keine großen Abweichungen erkennbar sind.

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 3.1 %, dies liegt für 4 Durchläufe im Rahmen bedeutet aber auch, dass dies beim Einrichten einer automatischen Pipeline berücksichtigt werden sollte.

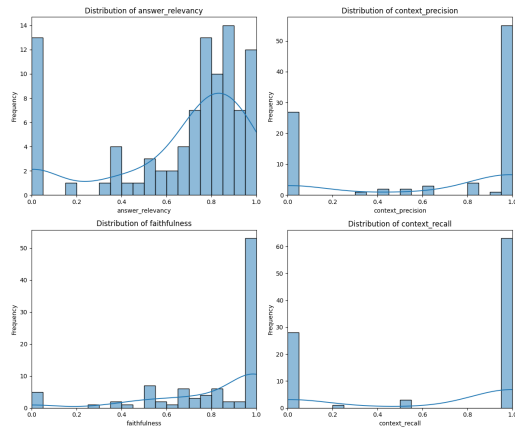


Abbildung 5.4: DeepSeek Ergebnis für 100 Fragen

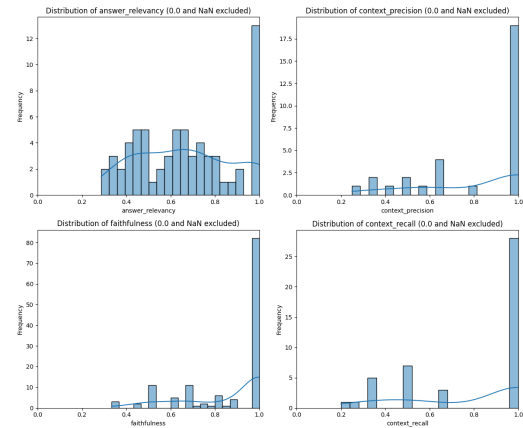


Abbildung 5.5: ChatGPT Ergebnis für 100 Fragen



Abbildung 5.6: Bewertung der vier Durchläufe mit DeepSeek

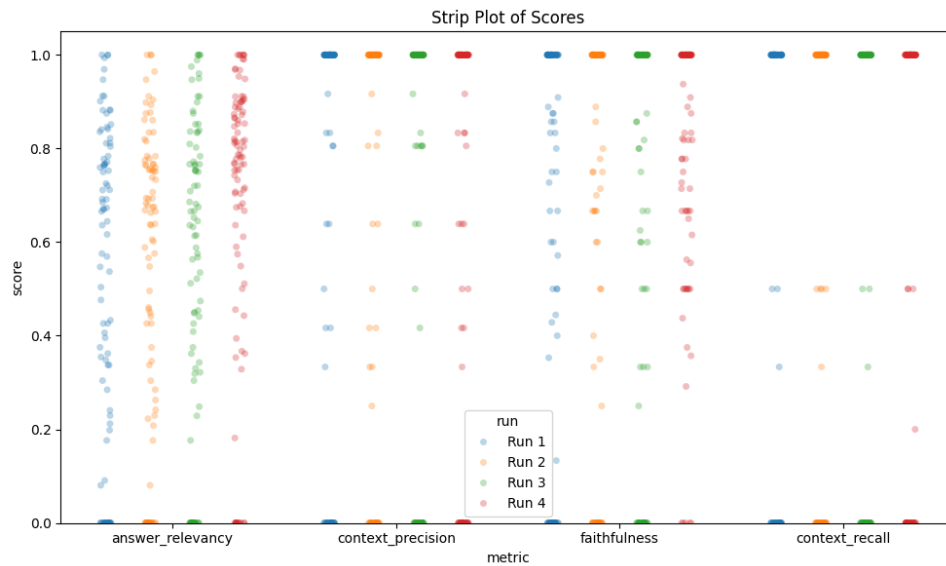


Abbildung 5.7: Bewertung der vier Durchläufe mit GPT-4

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 5.7: Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für DeepSeek

Metrik	Mean 1	Mean 2	Mean 3	Mean 4	Std 1	Std 2	Std 3	Std 4
Answer Relevancy	0.490	0.493	0.666	0.681	0.340	0.351	0.315	0.314
Faithfulness	0.632	0.613	0.815	0.800	0.434	0.444	0.273	0.295
Context Precision	0.669	0.684	0.666	0.670	0.445	0.445	0.444	0.445
Context Recall	0.663	0.667	0.681	0.671	0.461	0.466	0.458	0.456

Tabelle 5.8: Durchschnittswerte und Standardabweichungen der Metriken über vier Durchläufe für GPT-4

5.1.5 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 gesamt Bewertung als sehr stabil herausgestellt.

Answer Relevancy

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

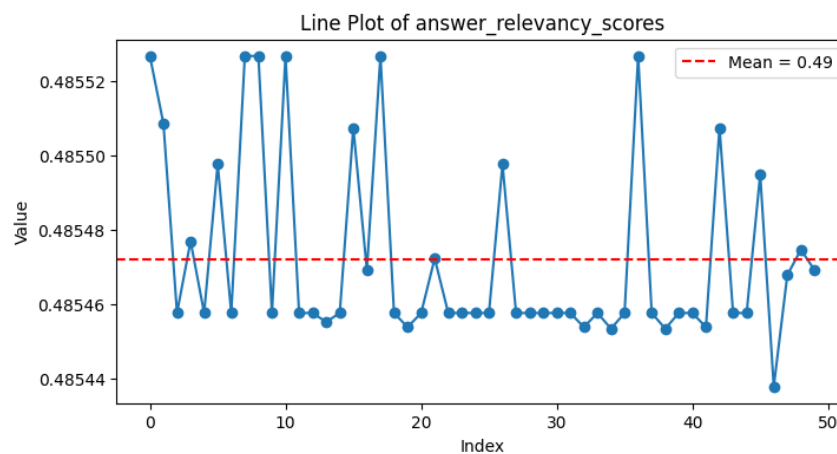


Abbildung 5.8: Abweichungen des Answer Relevancy Scores.

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.

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

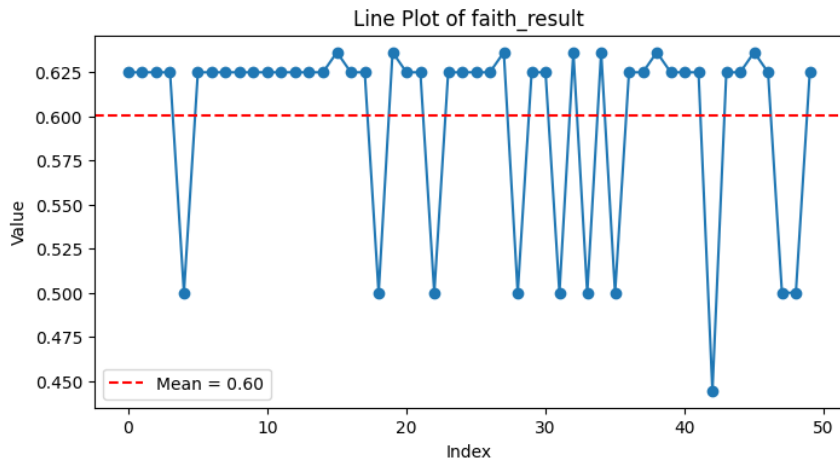


Abbildung 5.9: Abweichungen des Faithfulness Scores.

Ausführungszeiten DeepSeek

Da es bei OpenAi zu den Rate Limits kommen kann wurde die Anzahl an gleichzeitigen Abfragen von 16 auf 1 reduziert, da sonst besonders bei längeren durchläufen zu Problemen kommt. Das führt zu einer deutlichen verschlechterung der Ausführungszeit, bei dem Testset mit 15 Fragen sind wir von 2 Minuten mit maximal 16 gleichzeitigen Anfragen auf 7 Minuten bei maximal einer gleichzeitigen Anfrage. 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 5.9: Dauer der Evaluation pro Dokumentenzahl mit OpenAI

Ausführungszeiten OpenAi

Für die Bewertung des Testset mit 300 Fragen (400 Dokumente) wurde Tracing genutzt, die lässt uns genauer gucken, 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_recal und eine für faithfullnes.

Mit den 20 fehlgeschlagenen Metriken kommen wir auf eine Fehlerrate 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 5.10: Dauer der Evaluation pro Dokumentenzahl

5.1.6 Kostenberechnung

Die Bewertung des RAGs, mit den 300 Fragen, 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: 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): $720 \times 17\text{ h} = 12.240\text{ h} \approx \mathbf{1\text{ Jahr, 4 Monate, 10 Tage}}$

5.2 Abhängigkeit der Metriken untereinander

Dadurch, dass die Metriken für den Kontext einen früheren Schritt in der Abfrage an ein RAG bewerten als die für die Antwort, ergeben sich gewisse Abhängigkeiten.

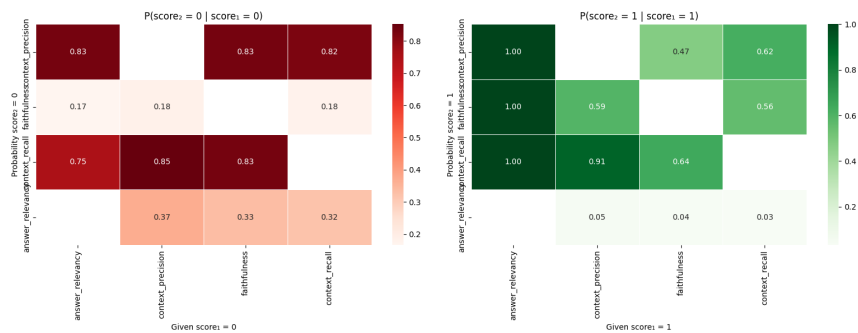


Abbildung 5.10: Abweichungen des Answere Relevancy Scores.

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 von dem OpenAI Rag für die 300 Fragen gezeigt, es lässt sich sehen, die faithfulness deutlich weniger von

5.3 Identifikation von Interessenten

6 Zusammenfassungen

6.1 Benutzung von RAGS

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

6.2 Testsets

Die generierung von Testsets ist eines der Alleinstellungsmerkmale von RAGAs, die generierung der Testsets hat sich aus Softwaretechnischer Sicht als unkompliziert herausgestellt. Es gab zu den Wichtigen Themen ausreichend Dokumentation und Beispiele.

Es ist mit RAGAs möglich Testsets zu generieren, jedoch gibt es mehrere Faktoren die Qualität und Praxistauglichkeit beeinflussen.

- Die fähigkeit des LLMs zuverlässig hochwertige antworten zu generieren
- Die Dokumente, desto komplexer und zusammenhangslos die Dokumente sind desto schlechter lassen sich Fragen generieren

Bei der Generierung von Testsets mit DeepSeek kam es alleine 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 Testsets mit OpenAIs GPT-4.1 hatte in Bezug auf nicht generierte oder Fragen zu irrelevanten Themen eine deutlich niedrigere Fehlerquote. Es gibt einen Ausreißer mit 205, der Rest bleibt jedoch deutlich unter 10 %. Die manuelle Auswertung hat hier aber auch gezeigt, dass es viele Fragen, bis zu 60 % bei ungefilterten Dokumenten Mängel aufweisen. Bei gefilterten Dokumenten kommt GPT-4.1 auf durchschnittlich 17,5 % und halbiert damit die Fehlerquote im Vergleich mit DeepSeek.

Die Qualität des Testsets ist entscheidend da sich hier entstanden Fehler weiter bis in die Bewertung durchziehen und eine korrekte Bewertung des eigentlichen RAGs verzerren!

Die bei den Versuchen generierte Testsets lassen Zweifel an einer zuverlässigen und hochwertigen generierung von Fragen aufkommen.

6.3 Bewertung

Auch das Generieren von Bewertungen hat sich mithilfe von RAGAS als einfach umzusetzen 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-4.1 30 % der Fragen nicht richtig beantwortet wurden. Dies lässt sich Teils auf die ungültigen Fragen in den Testsets zurückführen.

Bei den Metriken lässt sich sagen, dass die Metriken zum Kontext (recall und precision) gut abschneiden. Die faithfulness zeigt eine erhöhte Abweichung zu der menschlichen einschätzung und sollte mit run 10 %

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

6.4 Fazit

Insgesamt ist RAGAS kein kompletter Ersatz für die Menschliche Bewertung von RAGs. Die Idee hinter RAGAS Fragen ohne menschliches zu tun zu generieren um Zeit zu ersparen ist mit besseren LLMs teilweise gelungen. Um jedoch ein aussagen kräftiges und zuverlässiges Ergebniss zu generieren ist eine menschliche Kontrolle an mehreren Stellen notwendig. Zuerst bei der Auswahl der Dokumente, hier muss sowohl ein Verständniss vorhanden sein, wie gut LLMs mit welchen Daten umgehen können als auch welche Daten für das KMU relevant sind. Nach der generierung der Testsets sollte erneut ein Mensch über die Fragen gucken um grob Falsche Fragen zumindestens 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 genauer der Anwendungsfall betrachtet werden.

6.5 Zukunftsausblick

Für Unternehmen bieten LLMs und RAGs großes Potential für Kostenersparnisse, die Qualitätskontrolle spielt dabei eine immer größere Rolle. RAGAs bieten gute Ansätze um die Qualitätskontrolle zu automatisieren, dass RAGAs in Zukunft in die Prozesse zur Bewertung solcher Systeme einfließen ist daher sehr wahrscheinlich.

6.6 Reflektieren der Arbeit

Literatur

- [1] ExplodingGradients. *Integrations – How-to guide*. Zugegriffen am 18. Juni 2025. Mai 2025. URL: <https://docs.ragas.io/en/latest/howtos/integrations/>.
- [2] 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/>.
- [3] 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.
- [4] 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/>.
- [5] 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>.
- [6] 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>.
- [7] Ragas. *Context Entities Recall*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_entities_recall/.
- [8] Ragas. *Context Precision*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_precision/.
- [9] Ragas. *Context Recall*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/context_recall/.

- [10] Ragas. *Faithfulness*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/faithfulness/.
- [11] Ragas. *Noise Sensitivity*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/noise_sensitivity/.
- [12] Ragas. *Nvidia Metrics*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/nvidia_metrics/.
- [13] Ragas. *Query types in RAG*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/test_data_generation/rag/#query-types-in-rag.
- [14] Ragas. *Response Relevancy*. Accessed: 2024. 2024. URL: https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/answer_relevance/.
- [15] 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>.
- [16] Wikipedia. *Confusion matrix*. Accessed: 2024. 2024. URL: https://en.wikipedia.org/wiki/Confusion_matrix.
- [17] 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