Explorative Untersuchung der Effektivität semantischer Klassifizierung in der Abrufung für RAG-Systeme

Bachelorarbeit

vorgelegt am 06. Mai 2024

Fakultät Wirtschaft

Studiengang Wirtschaftsinformatik – Data Science

Kurs WWI2021F

von

Tien Dee Lin

Betreuer in der Ausbildungsstätte:

IBM Research

Cesar Berrospi Ramis

Senior Research Scientist

DHBW Stuttgart:

Prof. Dr., Kai Holzweißig

Inhaltsverzeichnis

Abkürzungsverzeichnis [1]

ANN	=	Approximate Nearest Neighbor
AP	=	Average Precision
BART	=	Bidirectional and Auto-Regressive Transformers
BERT	=	Bidirectional Encoder Representations from Transformers
BM-25	=	Best Matching 25
DARPA	=	Defense Advanced Research Projects Agency
DPR	=	Dense Passage Retriever
DSR	=	Design Science Research
EM	=	Exact Match
GLUE	=	General Language Understanding Evaluation
GPT-4	=	Generative Pre-trained Transformer 4
GQM	=	Goal Question Metric
IR	=	Information Retrieval
LLM	=	Large Language Model
MAP	=	Mean Average Precision
MRR	=	Mean Reciprocal Rank
MVP	=	Minimum Viable Product
NDCG	=	Normalized Discounted Cumulative Gain
NIST	=	National Institute of Standards and Technology
NLP	=	Natural Language Processing
ODQA	=	Open Domain Question Answering
ORQA	=	OpenRetrieval Question Answering System
QNLI	=	Question-answering Natural Language Inference
RAG	=	Retrieval Augmented Generation
TF-IDF	=	Term Frequency-Inverse Document Frequency
TREC	=	Text Retrieval Conference

[1] Dieses Abkürzungsverzeichnis wurde am 04.05.2024 mithilfe von GPT-4 in alphabetischer Reihenfolge sortiert

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Motivation

Mit der Einführung der generativen Sprachmodelle GPT-3.5 und GPT-4 wurde die disruptive Kraft dieser Technologie in Industrie, Wirtschaft und Gesellschaft deutlich.[2] Nach der anfänglichen Begeisterung, die eine Zukunft versprach, in der generative KI eine bedeutende Rolle im Alltag und in Geschäftsprozessen spielt, folgte eine Phase der Ernüchterung – insbesondere durch das Problem der Halluzination.[3] In der Wissenschaft wird Halluzination in generativen Sprachmodellen als die Erzeugung von Inhalten definiert, die faktisch unkorrekt oder irreführend sind. Dies tritt auf, wenn das Modell Texte generiert, die veraltete, fehlerhafte, oder frei erfundene Informationen enthalten.[4] Die Bekämpfung dieses Halluzinationsrisikos stellt eine der zentralen Herausforderungen in der Forschung zu generativen Sprachmodellen dar, da viele kritische Anwendungsbereiche wie das Gesundheitswesen, der Journalismus oder das Rechtswesen auf absolute Genauigkeit angewiesen sind und selbst kleinste Fehler schwerwiegende Folgen haben können.[5] Zudem beschränkt das Halluzinationsrisiko auch den potenziellen Mehrwert und die Produktivitätssteigerung durch den Einsatz dieser Modelle in weniger kritischen Bereichen, da der generierte Inhalt immer manuell überprüft werden muss.

Ein vielversprechendes Konzept zur Lösung dieses Problems ist das „Retrieval Augmented Generation“ (RAG).[6] Diese Technologie nutzt Informationen aus einem externen Speicher und ruft, basierend auf der Nutzeranfrage, die korrekten Informationschunks ab. Die abgerufenen textuellen Informationen dienen dem antwortgenerierenden Sprachmodell als Kontext, um faktisch korrekte Antworten bereitzustellen. Dieses Verfahren ermöglicht es, aktuelle Informationen aus einem dynamisch aktualisierbaren Speicher zu nutzen, anstatt sich auf möglicherweise veraltete Trainingsdaten des Sprachmodells zu verlassen.[7] Dadurch können Antworten kontinuierlich an den neuesten Informationsstand angepasst und erweitert werden, ohne dass ein aufwendiges Neu-Trainieren oder Feinabstimmung des generativen Modells erforderlich ist.[8]

Problemstellung

Ein wesentlicher Bestandteil des RAG-Systems ist das Abrufungssystem, das „R“ in RAG. Dieses Thema, von etablierten Konzepten der Informationsabrufung vergangener Generationen bis zu aktuellen Abrufungsmethoden für RAG, wird in der Wissenschaft intensiv diskutiert.[9]

---

[2] Vgl. Wong 2024, S. 1 ff.
[3] Vgl. Yuyan Chen et al. 2023, S. 245 f.
[4] Vgl. Rawte, Sheth, Das 2023, S. 1
[5] Vgl. Rawte, Sheth, Das 2023, S. 1 f.
[6] Vgl. Béchard, Ayala 2024, S. 1 f.
[7] Vgl. Lewis et al. 2020, S. 1 f.
[8] Vgl. Mosbach et al. 2023, S. 12291 ff.
[9] Vgl. Gao et al. 2024, S. 4 ff.

Trotz der zunehmenden Bedeutung von Diskussionen um RAG, wo stetig neue Abrufungsmethoden erörtert werden, ist eines offensichtlich: Die aktuellen Systeme sind nicht perfekt und bieten insbesondere bei der Abrufungsgenauigkeit Verbesserungspotenzial.[10] Kein Abrufungssystem kann derzeit behaupten, immer zuverlässig die korrekte Informationspassage aus einem umfangreichen Dokumentenkorpus zu identifizieren. Daher wird im wissenschaftlichen Diskurs kontinuierlich mit neuen Systemen experimentiert und deren Effektivität bewertet. In der Praxis variieren die Herausforderungen je nach Anwendungsfall erheblich. Unternehmen benötigen oft spezifische Abrufungsdaten, die auf ihre Geschäftsprozesse zugeschnitten sind, und es stehen oftmals keine repräsentativen, wissenschaftlich validierten Evaluierungsdaten zur Verfügung.[11] Zudem ist es nicht immer wirtschaftlich, allgemeine Abrufungssysteme aufwendig auf domänenspezifische Daten abzustimmen. Aus diesen Gründen ist es relevant zu untersuchen, ob Möglichkeiten existieren, die Abrufungsgenauigkeit für RAG-Systeme mit geringerem Entwicklungs- und Ressourcenaufwand zu verbessern.

Zielsetzung

Das Ziel dieser Arbeit ist es, explorativ zu untersuchen, ob die Abrufungsgenauigkeit durch die Modifikation des bestehenden Abrufungssystems von IBM Deepsearch verbessert werden kann, indem eine semantische Klassifizierung von Fragen und Passagen integriert wird. IBM Deepsearch ist ein RAG-System, das von IBM Research entwickelt wurde und dessen Abrufungskomponente dem aktuellen Stand der Forschung entspricht. Die semantische Klassifizierung wird in diesem Kontext als ressourceneffizient betrachtet, da sie mittels effizienter Methoden wie beispielsweise der „Support Vector Machine“ durchgeführt werden kann.[12] Weiterhin wird angestrebt, die Forschungsarbeit so zu gestalten, dass die Prozesse weitgehend reproduzierbar und die Ergebnisse möglichst generalisierbar sind.

Forschungsmethodik

In der vorliegenden Arbeit wird das Goal-Question-Metric (GQM) Modell nach Basili, Caldiera und Rombach (1994) angewandt, um eine präzise und systematische Zieldefinition durchzuführen. Basierend auf dem übergeordneten Ziel auf konzeptioneller Ebene werden Unterfragen formuliert, die mittels spezifischer Metriken untersucht und beantwortet werden. Hierzu werden quantitative Metriken eingesetzt, um eine vergleichbare Messung der Abrufungsgenauigkeit zu ermöglichen, wobei die Auswahl und Begründung der Metriken an geeigneter Stelle dargelegt wird. Die Untersuchung der Unterfragen auf operativer Ebene erfolgt gemäß der Design Science Research (DSR) Methodik nach Pfeffers et al. (2007). Diese wurde gewählt, da der DSR-Prozess iterative Lösungsansätze ermöglicht, die separat

[10] Vgl. Siriwardhana et al. 2023, S. 8
[11] Vgl. Ji Ma et al. 2021, S. 3
[12] Vgl. Zhang, Lee 2003, S. 28

implementiert und evaluiert werden können. Dies erlaubt es, den Forschungsfortschritt nach jeder Iteration präzise zu erfassen und gegebenenfalls eine neue Iteration zur Verbesserung oder Anpassung der Lösungsansätze zu initiieren. Die iterative Vorgehensweise des DSR unterstützt insbesondere die Implementierung von Klassifikationssystemen für semantische Kategorisierungen in der Abrufung, welche die Daten des Ziel-Datensatzes mit angemessener Genauigkeit klassifizieren müssen. Auch ist diese Metrik vorteilhaft, da sie es ermöglicht, in den Iterationen getrennt verschiedene Methoden der Verwendung von Klassifizierungen sowie verschiedene Klassifikationstaxonomien auszuprobieren und zu evaluieren.

Aufbau der Arbeit

Kapitel 1 bietet eine Einführung in die Thematik, leitet die Problemstellung aus relevanter Forschungsliteratur ab und legt die Zielsetzung sowie die verwendete Methodik dieser Arbeit dar. Kapitel 2 dient der Erörterung des aktuellen Forschungsstands, beginnend mit einer Einführung in die Thematik von RAG. Es folgt die Darlegung des aktuellen Stands der Forschung und Diskussion der relevanten Konzepte, die durch eine strukturierte Literaturrecherche im Themenbereich der Abrufung für RAG identifiziert wurden. Kapitel 3 verfeinert die Zielsetzung, präsentiert und begründet das gewählte Forschungsdesign. Kapitel 4 behandelt den praktischen Teil der Arbeit, präsentiert die Ergebnisse der jeweiligen Iterationen und erläutert besonders relevante Aspekte der Implementierung. Kapitel 5 diskutiert die Ergebnisse aus Kapitel 4 und führt sie zu Schlussfolgerungen zusammen. Kapitel 6 bildet den Abschluss dieser Arbeit, bietet eine umfangreiche kritische Reflexion der Ergebnisse und Methodik, erörtert, ob der Auftrag der Arbeit erfüllt wurde und diskutiert die Implikationen dieser Arbeit für Theorie und Praxis. Abschließend wird ein Ausblick auf potenzielle weiterführende Forschung gegeben.

2 Diskussion des aktuellen Stands der Forschung und Praxis

In diesem Kapitel werden die theoretischen Grundlagen und Hintergründe im Forschungsbereich des RAG für Fragebeantwortung, mit einem besonderen Fokus auf das Abrufungssystem, erörtert, um ein vertieftes Verständnis der zugrundeliegenden Arbeit zu bieten. Das Kapitel ist in vier Unterkapitel gegliedert. Kapitel 2.1 führt das allgemeine Konzept von RAG ein und bietet einen spezifischen Überblick über die für diese Arbeit relevanten Aspekte des RAG-Systems. Kapitel 2.2 konzentriert sich auf die Nutzung von RAG-Systemen zur Fragebeantwortung. Kapitel 2.3 widmet sich dem Abrufungssystem, dem „R" in RAG, wobei eine Konzeptmatrix nach Webster und Watson (2002) erstellt (Unterkapitel 2.3.5) und nach dieser Konzeptmatrix relevante Konzepte erläutert und diskutiert werden. Dieses Kapitel ist weiter in fünf Unterkapitel unterteilt, die sich den übergeordneten Themenbereichen wie Abrufungsmethoden, Evaluationsmetriken sowie der synthetischen Generierung von Trainings- und Evaluationsdaten widmen. Die Konzepte werden im letzten Unterkapitel (2.3.5) speziell hinsichtlich ihrer praktischen Implementierung diskutiert. In Kapitel 2.4 werden relevante Konzepte zur semantischen Klassifizierung von Fragen vorgestellt und ihre Relevanz für den Informationsabrufung erörtert.

2.1 Einführung in Retrieval Augmented Generation

Retrieval Augmented Generation ist eine Methodik, die konventionelle Generierungsmodelle durch die Integration externer Datenquellen bereichert, um die Genauigkeit und Relevanz der erzeugten Ergebnisse signifikant zu verbessern.[13] Dies wird ermöglicht, indem sichergestellt wird, dass die generierten Inhalte auf akkurat abgerufenen relevanten Informationen basieren.[14]

Erstmals vorgestellt von Lewis et al. im Jahr 2023, adressiert RAG gezielt die Herausforderungen der Halluzinationen und der Wissensinflexibilität, welche bei umfangreichen Sprachmodellen auftreten.[15] Obwohl diese Modelle in der Lage sind, ihre umfassenden Trainingsdaten als eine parametrisierte Wissensbasis zur Generierung von Ergebnissen zu nutzen,[16] offenbaren sie bestimmte Einschränkungen: Die Erweiterung oder Modifikation ihrer Informationsbasis ist mit erheblichem Aufwand verbunden,[17] direkte Einblicke in die Denkprozesse hinter ihren Ergebnissen sind nicht direkt möglich,[18] und sie tendieren dazu, irreführende oder erfundene Inhalte (Halluzination) zu produzieren.[19] Die Integration von akkuraten Informationen aus externen Datenbanken über RAG verstärkt signifikant die Genauigkeit und Glaubwürdigkeit

[13] Vgl. Gao et al. 2024, S. 1
[14] Vgl. Siriwardhana et al. 2023, S. 1
[15] Vgl. Lewis et al. 2020, S. 1 f.
[16] Vgl. Roberts, Raffel, Shazeer 2020, S. 2 ff.
[17] Vgl. Modarressi et al. 2023, S. 1 ff.
[18] Vgl. Lewis et al. 2020, S. 1
[19] Vgl. Zhang et al. 2023, S. 3

der generativen Modelle, speziell für Anwendungsfälle, wo domänenspezifisches Wissen benötigt wird.[20] Dies erlaubt es auch, dass Wissen kontinuierlich auf dem neuesten Stand gehalten werden kann und domänenspezifische Informationen integriert werden können.[21]

Neben dem etablierten Einsatzbereich in der Textgenerierung,[22] die primär auf textuellen Informationen basiert, bietet RAG das Potenzial, multimodale Informationen abzurufen[23] und in der Generierung von beispielsweise Bildern Anwendung zu finden.[24] Die vorliegende Arbeit konzentriert sich jedoch ausschließlich auf den Einsatz von RAG im Kontext der Generierung textueller Inhalte zur Beantwortung von Fragen, bei dem nur textuelle Informationen abgerufen werden.

Im Folgenden wird ein Schema einer "naiven" RAG-Anwendung gemäß dem Retrieve-Read-Framework präsentiert, ohne Modifikation des Anwenderinputs vor dem Retrieval.[25]

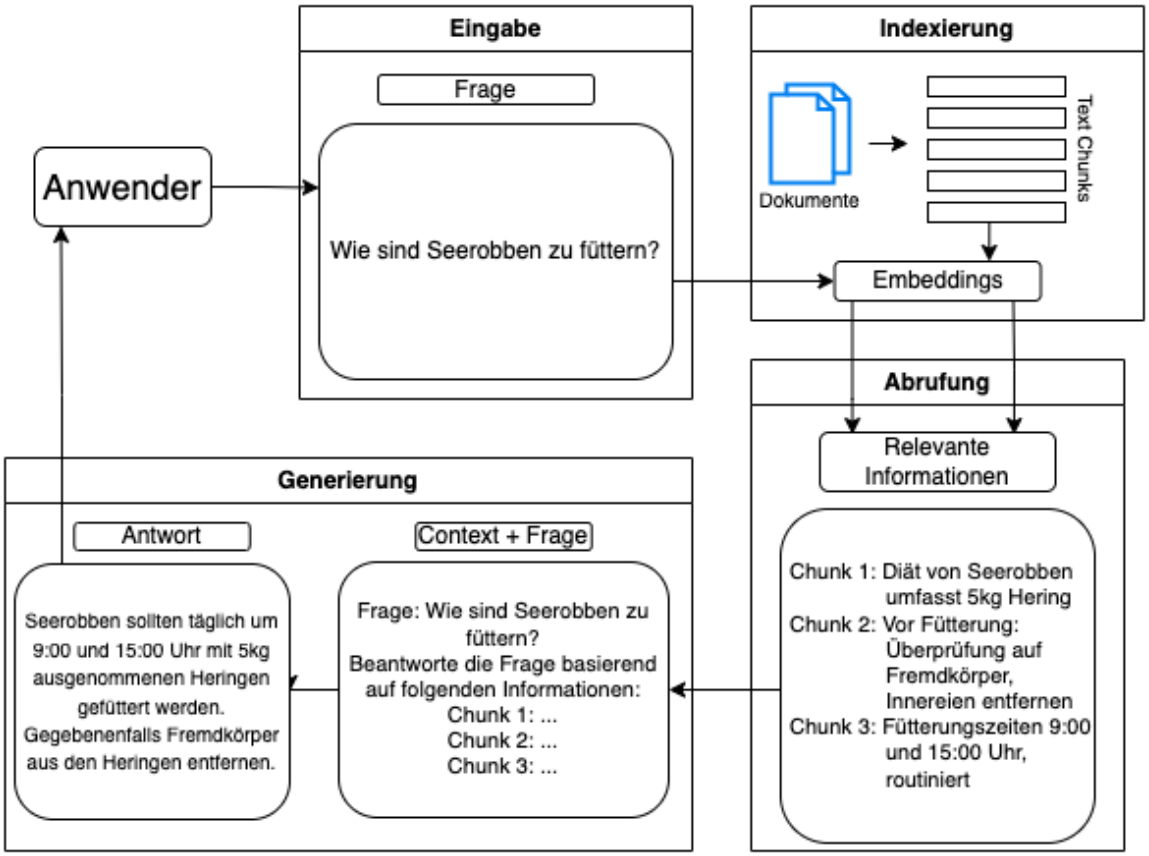


Abb. 1: Schematische Darstellung eines RAG-Systems[26]

[20] Vgl. Siriwardhana et al. 2023, S. 1
[21] Vgl. Gao et al. 2024, S. 1
[22] Vgl. Li et al. 2022, S. 1 ff.
[23] Vgl. Zhao et al. 2023, S. 2 f.
[24] Vgl. Blattmann et al. 2022, S. 3 ff.
[25] Vgl. Ma et al. 2023, S. 3
[26] nach Gao et al 2024, mit Veränderungen

Unter "Naivem RAG" versteht man die grundlegendste Methodik der Retrieval Augmented Generation, welche die Schritte der Indexierung, des Abrufs und der Generierung umfasst.[27] Zuerst wird der Indexierungsschritt durchgeführt, bei dem die Daten oder externes Wissen, welche später über RAG abgerufen werden sollen, aus ihren originären Formaten wie zum Beispiel PDF[28] oder HTML-Seiten[29] extrahiert werden und aufbereitet werden. Die extrahierten Informationen werden anschließend als Text gespeichert, welche dann in kleinere Abschnitte, sogenannte 'Chunks', unterteilt werden.[30] Diese Unterteilung erfolgt aus zweierlei Gründen: zum einen, um die Anpassung an die begrenzte Kontextlänge des Sprachmodells im Generierungs-Schritt zu gewährleisten,[31] und zum anderen, um eine zielgerichtete Abfrage zu ermöglichen, indem genau der Chunk (oder auch „Passage“) mit der spezifisch gesuchten Information abgerufen wird.[32]

Die segmentierten Chunks/Passagen werden anschließend durch ein Encoder-Modell in kontextualisierte, dichte Vektoren, sogenannte Dense Embeddings, umgewandelt.[33] Diese Vektoren bilden die Indeces der originalen Passagen, sodass sie als Suchindex für die Abfrufung fungieren.[34]

Im Abrufungsschritt richtet der Anwender eine Frage an das System welches ebenfalls mit einem Encoder in einen entsprechend kontextualisierten Dense-Embedding-Vektor transformiert wird.[35] Die semantische Nähe zwischen der Anfrage und den Chunks wird beispielsweise über das Kosinusähnlichkeitsmaß oder das Skalarprodukt zwischen dem Embedding-Vektor der Frage und den Embedding-Vektoren (Indices) der Chunks ermittelt.[36] Dadurch können jene Chunks identifiziert werden, die die höchste Übereinstimmung mit der Frage aufweisen und somit wahrscheinlich die relevanten Informationen oder Kontext zur Frage enthalten.[37] Zusätzlich zu dieser Methode des Abrufens gibt es auch andere Ansätze, die spärliche Vektoren (engl. "sparse vectors") verwenden und auf lexikalischen Algorithmen wie BM-25[38] oder TF-IDF[39] basieren, um relevante Passagen zu identifizieren.

Im Generierungsschritt werden die abgerufenen Chunks zusammen mit der Anfrage, beide in Textform, und unter Verwendung einer Prompt-Struktur (siehe Abb. 1) einem generativen

[27] Vgl. Gao et al. 2024, S. 3
[28] Vgl. Livathinos et al. 2021, S. 15137
[29] Vgl. Breuel 2003, S. 11
[30] Vgl. Ma et al. 2021, S. 5
[31] Vgl. Gao et al. 2024, S. 3
[32] Vgl. Karpukhin et al. 2020, S. 2
[33] Vgl. Reichman, Heck 2024, S. 1
[34] Vgl. Gao et al. 2024, S. 3
[35] Vgl. Karpukhin, Oğuz, et al. 2020, S. 3
[36] Vgl. Mitra, Craswell 2017, S. 23
[37] Vgl. Gao et al. 2024, S. 4
[38] Vgl. Sarrouti, Ouatik El Alaoui 2017, S. 100
[39] Vgl. Neves 2015, S. 4

Sprachmodell, wie zum Beispiel GPT-4[40] oder BART[41], übergeben. Das Modell generiert basierend auf diesem Prompt die Antwort für die Anfrage des Anwenders.[42] Über dieser Implementierung von RAG hinaus existieren zahlreiche komplexere Varianten und Verfeinerungen dieser Technik, die in den nachfolgenden Kapiteln detailliert betrachtet und diskutiert werden.

Neben dem klassischen Anwendungsfall von RAG für Dialog-/Chat-Systeme zur Optimierung der Korrektheit und Genauigkeit der Antworten, findet RAG auch Anwendungen in vielen anderen Bereichen[43], wie zum Beispiel beim Training von Sprachmodellen, indem fehlerhafte Trainingsdaten durch RAG "repariert“ werden bzw. fehlende Informationen ergänzt werden.[44] Auch kann die Zusammenfassung von Dokumenten durch Sprachmodelle optimiert werden, indem vor der eigentlichen Zusammenfassung, basierend auf der Anfrage, eine Vorlage der gewünschten Zusammenfassung abgerufen wird und die Zusammenfassung dann unter Berücksichtigung dieser Vorlage umgesetzt wird.[45] Die vorliegende Arbeit fokussiert sich auf den spezifischen Anwendungsfall der Fragebeantwortung (engl. Question-Answering) durch RAG.

2.2 RAG für Fragebeantwortung und Domänadaption

Open-Domain Question Answering (ODQA)[46] stellt den klassischen Einsatzbereich für Retrieval Augmented Generation (RAG) dar, bei dem, wie bereits erörtert, ein generatives Sprachmodell die Nutzerfragen auf Basis von abgerufenen externen Daten beantwortet. Nach derzeitigem Forschungsstand sind die meisten RAG-Implementierungen für die domänenoffene Fragebeantwortung prädestiniert, da RAG in wissenschaftlichen Arbeiten vorwiegend mit auf Wikipedia basierenden Datensätzen trainiert und evaluiert wurde.[47] Ein weiterer Faktor, der die Eignung von RAG-Implementierungen für die Beantwortung von domänenübergreifenden Fragen unterstützt, ist die weitverbreitete Verwendung des auf Wikipedia-Daten trainierten Dense Passage Retrievers (DPR) als Komponente für die Informationsabrufung, wobei dieser Aspekt auch auf andere Retriever zutrifft.[48] Die domänenspezifische Fragebeantwortung mit RAG, etwa das Beantworten von Fachfragen aus wissenschaftlichen Publikationen oder medizinischen Dokumenten, kennzeichnet jedoch eine signifikante Forschungslücke.[49]

---

[40] Vgl. Unlu et al. 2024, S. 7 f.
[41] Vgl. Siriwardhana et al. 2023, S. 4
[42] Vgl. Gao et al. 2024, S. 4
[43] Vgl. Li et al. 2022a, S. 7
[44] Vgl. Guu et al. 2020, S. 3
[45] Vgl. Peng et al. 2019, S. 2555 ff.
[46] Vgl. Zhu et al. 2021, S. 6 ff.
[47] Vgl. Siriwardhana et al. 2023
[48] Vgl. Karpukhin, Oğuz, et al. 2020, S. 4 ff.
[49] Vgl. Siriwardhana et al. 2023, S. 2

Die Adaptation an spezifische Domänen ist im Feld des Natural Language Processing (NLP) ein weitverbreitetes und intensiv erforschtes Problem.[50] Im Zeitalter des Pre-train-Finetune-Paradigmas hat diese Anpassung zusätzlich an Bedeutung gewonnen, da dieser Ansatz die Weiterentwicklung von generalistischen zu spezialisierten Sprachmodellen beschreibt.[51] Der Ansatz, individuelle neue Sprachmodelle für jede Domäne feinabzustimmen oder sie spezifisch auf domänenspezifischen Datensätzen zu trainieren, erweist sich jedoch nicht stets als optimal, angesichts der damit verbundenen hohen Kosten.[52] Aus diesem Grund spielt die Domänenadaption von RAG eine wesentliche Rolle: Sie kann potenziell die Notwendigkeit separater Modelle für jede Domäne eliminieren, indem sie ein einzelnes generatives Modell nutzt, welches Wissen verschiedener Domäne verwendet und somit die Kosten reduziert. Darüber hinaus führen Fortschritte in der domänenspezifischen Anpassungsfähigkeit zu einer Verbesserung der allgemeinen Leistungsfähigkeit bei ODQA.[53]

2.3 Konzepte von RAG für Fragebeantwortung

Der Schwerpunkt dieser Arbeit liegt auf dem Abrufungssystem eines RAG-Systems. In den nachfolgenden Unterkapiteln wird ausgehend von der Konzeptmatrix in Kapitel 2.3.1 eine Diskussion und Erläuterung der relevanten Konzepte dieses Schwerpunktes vorgenommen. Ferner werden die entscheidenden Aspekte der semantischen Klassifizierung von Fragen und Antworten beleuchtet, die einen zentralen Forschungsbeitrag dieser Arbeit darstellen.

2.3.1 Konzeptmatrix

Tabelle 1 zeigt die Konzeptmatrix nach Webster und Watson (2002), die speziell auf die Abrufungssysteme im Kontext von RAG für die Fragebeantwortung ausgerichtet ist. Die Konzeptmatrix wird auf der Grundlage einer umfangreichen Literaturrecherche erstellt. In dieser Recherche wird die gesichtete Literatur analysiert, während die Konzeptmatrix die Inhalte dieser Literatur um relevante Konzepte gruppiert.[54] Die Recherche erfolgte gemäß einer empfohlenen Methodik, beginnend mit der Durchsicht führender und für ihre Qualität anerkannter wissenschaftlicher Journale und Konferenzen zur Identifizierung relevanter Artikel. Nach deren Bearbeitung fand eine Rückwärtssuche statt, um frühere zitierte Arbeiten zu überprüfen, gefolgt von einer Vorwärtssuche, um Werke zu finden, die diese Artikel zitierten.[55] Diese methodische Literaturrecherche, ergänzt durch die Konzeptmatrix, zielt darauf ab, eine vollständige Übersicht zu bieten, die sich nicht auf eine bestimmte Auswahl von Zeitschriften, Autoren oder geografischen Regionen beschränkt und ist strikt konzeptorientiert.[56]

---

[50] Vgl. Ramponi, Plank 2020, S. 1
[51] Vgl. Guo, Yu 2022, S. 4 f.
[52] Vgl. Xu et al. 2023, S. 56
[53] Vgl. Siriwardhana et al. 2023, S. 2
[54] Vgl. Webster, Watson 2002, S. xvii
[55] Vgl. Webster, Watson 2002, S. xvi
[56] Vgl. Webster, Watson 2002, S. xv

Autor	Konzepte																		
	Retriever								Evaluationsmetrik							Weiteres			
	Dense Passage Retriever	andere Dense Retriever	BM-25	TF-IDF	Hybrid	Nearest Neighbor Search	Maximum Inner Product Search	Cross-Attention-Rescorer	Precision (Mean, Average)	F1	nDCG	Mean Reciprocal Rank	Accuracy (Top-k)	Exact Match	Recall	Offene Domäne	Spezifische Domäne	Synthetisch generierte Daten	Synthetisch angereicherte Anfragen
Alawwad et al. 2024	x												x				x		
Soudani et al. 2024	x		x		x								x			x		x	
Kharitonova et al. 2024	x																x		
Yubo Wang et al. 2024	x				x	x		x					x				x		x
Levonian et al. 2023		x								x							x		
Zhang et al. 2023		x							x							x		x	
Wang et al. 2023	x												x	x	x	x			
Sirwardhana et al. 2023	x	x								x			x	x			x	x	
Pan et al. 2022	x								x	x		x		x	x	x			
Izacard et al. 2022	x					x		x				x			x	x			
Mao et al. 2021	x		x										x	x					x
Wang et al. 2021		x	x		x			x	x		x	x			x	x			
Ma et al. 2021					x	x		x	x		x		x				x	x	
Lewis et al. 2021	x						x	x						x	x	x			
Karpukhin et al. 2021	x				x			x					x			x			
Lee et al. 2019		x					x							x		x			
Sarrouti et al. 2017			x						x	x					x		x		
Robertson et al. 2009			x						x		x	x				x			
Ramos 2003				x									x			x			

Tab. 1: Konzeptmatrix zu RAG für textuelle Fragebeantwortung mit Fokus auf Retriever

2.3.2 Retriever

Ein Retriever, auch oft Information Retrieval (IR) System genannt, wird definiert als ein System, welches das Ziel hat, zu einer gegebenen Frage in natürlicher Sprache die relevantesten Dokumente oder Passagen abzurufen bzw. diese nach ihrer Relevanz zu ordnen. (engl. „ranking“).[57] Der Retriever, der einen zentralen Bestandteil des in Kapitel 2.1 beschriebenen naiven RAG-Ansatzes darstellt, ist für die Durchführung der Indexierungs- und Abrufungsschritte verantwortlich. Innerhalb dieses Kapitels wird eine eingehende Diskussion der verschiedenen Konzepte und Methodiken des Retrieval-Bereiches vorgenommen.

„Sparse“ Retriever

Ein "Sparse" Retriever identifiziert relevante Passagen basierend auf lexikalischen Eigenschaften im Verhältnis von Frage zur Passage, wie zum Beispiel die Häufigkeit spezifischer Wörter in der Frage und den Textpassagen.[58] Im Gegensatz zum "Dense" Retriever, sieht der "Sparse" Retriever Texte lediglich als eine Ansammlung von Wörtern (engl. "Bag-of-Words" representation), ohne den semantischen Zusammenhang zu berücksichtigen.[59]

TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) ist eine Methode, um die Relevanz eines Wortes (Term) zu berechnen, indem es die relative Häufigkeit von einem Wort in der Passage mit der inversen Häufigkeit des Wortes im Korpus (alle Passagen) betrachtet.[60]

Gegeben sei ein Korpus K, ein Term t und eine individuelle Passage (im Korpus) p mit $p \in K$.

Term Frequency (TF) ist die Häufigkeit eines Wortes in der bestimmten Passage und wird wie folgt berechnet:[61]

$$TF(t,p) = \frac{Anzahl\ von\ t}{Anzahl\ von\ Termen\ in\ p}$$

TF kann auch als absolute Häufigkeit des Terms in der Passage verwendet werden, hier wird eine Normalisierung vorgenommen, um den eventuellen Umstand auszugleichen, dass die Länge der Passagen nicht immer gleich ist und somit die Häufigkeit des Terms verzerrt ist.[62]

Inverse Document Frequency (IDF) ist die inverse Häufigkeit des Wortes im Korpus und wird wie folgt berechnet:[63]

---

[57] Vgl. Zhu et al. 2021, S. 5
[58] Vgl. Karpukhin, Oğuz, et al. 2020, S. 1
[59] Vgl. Wang, Zhuang, Zuccon 2021, S. 1
[60] Vgl. Ramos 2003, S. 2
[61] Vgl. Christian, Agus, Suhartono 2016, S. 289
[62] Vgl. Qaiser, Ali 2018, S. 25
[63] Vgl. Christian, Agus, Suhartono 2016, S. 289

$$IDF(t, K) = log\left(\frac{Gesamtzahl\ der\ Passagen\ p\ in\ K}{Anzahl\ der\ Passagen\ p\ mit\ Term\ t}\right)$$

Mit der Multiplikation von $TF \times IDF$ erhält man ein Maß für die Relevanz eines Terms t in Passage p, skaliert nach seiner Häufigkeit im Korpus. Der Term ist zuerst nur dann relevant, wenn nicht alle Passagen im Korpus diesen Term enthalten, da $\log\left(\frac{a}{a}\right) = log(1) = 0$. Der Term, oder seine Häufigkeit in p, wird relevanter, je weniger es Passagen im Korpus gibt, welche diesen Term enthalten (siehe Logarithmus).

Bei einer gegebenen Frage, die aus mehreren Wörtern besteht, wird für jedes Wort (in der Frage) der TF-IDF-Score in jeder Passage berechnet.[64] Durch die Summierung der berechneten TF-IDF Werte jeder Passage kann eine Rangfolge erstellt werden, welcher die Passagen nach ihrer Relevanz zur gestellten Frage ordnet.[65] Es können nun die Passagen mit der höchsten Summe dieser Werte zur Beantwortung der Frage herangezogen werden.[66]

BM-25

Best Matching 25 (BM-25) ist ebenfalls ein Ranking-Algorithmus, welcher zur Bewertung von Dokumenten oder Passagen (Passagen können als kurze Dokumente gesehen werden) nach ihrer Relevanz in Bezug auf eine bestimmte Frage konzipiert wurde.[67] Dieser Algorithmus, der in die Kategorie der „Sparse"-Retriever-Methoden fällt, baut auf den Grundlagen von TF-IDF auf, indem es Term-Frequenz (TF) und Inverse Dokument Frequenz (IDF) nutzt, führt jedoch bedeutende Verbesserungen ein, um die Effektivität des Rankings zu steigern.[68]

Zum einen implementiert BM-25 eine Sättigungsfunktion für die Term-Frequenz (TF), die den Anstieg der Passagenbewertung begrenzt, je häufiger der Term vorkommt.[69] Dieses Vorgehen korrigiert die bei TF-IDF entstehende unproportionale Score-Erhöhung, bei der eine Zunahme von einem auf zwei Vorkommen eines Terms den gleichen Effekt haben kann wie eine Zunahme von 100 auf 101.[70] Zum anderen berücksichtigt BM-25 die Länge der Passage in der Bewertung, indem es den Relevanzscore anhand der Abweichung von der durchschnittlichen Passagenlänge anpasst, wodurch verhindert wird, dass längere Texte einen verzerrenden Vorteil erhalten.[71] Im vorherigen Teil wurde TF-IDF mit einer Normalisierung von TF nach der

[64] Vgl. Ramos 2003, S. 2
[65] Vgl. Christian, Agus, Suhartono 2016, S. 289
[66] Vgl. Ramos 2003, S. 2
[67] Vgl. Rosa et al. 2021, S. 1 f.
[68] Vgl. Guo et al. 2020, S. 4
[69] Vgl. Whissell, Clarke 2011, S. 479 f.
[70] Vgl. Christian, Agus, Suhartono 2016, S. 288
[71] Vgl. Lv, Zhai 2011, S. 8

Passagenlänge vorgestellt. Jedoch wird TF auch nicht-normalisiert als absolute Häufigkeit verwendet, wobei eine Normalisierung in der Forschungsliteratur als „nützliche Heuristik“ betrachtet wird.[72]

Gegeben sei ein Korpus K, ein Term t und eine individuelle Passage (im Korpus) p mit $p \in K$. $tf(t,p)$ ist die absolute Häufigkeit von t in p, k_1 und b sind frei wählbare Parameter und θ ist das Verhältnis von der Passagenlänge der aktuellen Passage $|p|$ zur durchschnittlichen Passagenlänge im gesamten Korpus $avgpl$. IDF ist die bekannte Inverse-Dokument-Frequenz und q ist die Frage bestehend aus einer Menge von t. Der BM-25 Score lässt sich nun folgendermaßen berechnen:[73]

$$BM25(q,p) = \sum_{t \epsilon q \cap p} IDF\,(t) \times \frac{tf(t,p) \times (k_1 + 1)}{tf(t,p) + k_1 \times (1 - b + b \times \theta)}$$

$$\theta = \frac{|p|}{avgpl}$$

Bei einer gegebenen Passage wird der BM25 Score der Passage berechnet indem für jeden Term, welcher sowohl in der Frage und in der Passage vorkommt, der BM-25 Wert des Terms in der Passage berechnet wird und dann über alle Terme aufsummiert wird.[74] Man kann nun diese Berechnung für alle Passagen im Korpus unternehmen und bei gegebener Frage alle Passagen nach ihrer Relevanz zur Frage, ordnen und die relevantesten Passagen zur Beantwortung der Frage (im Generations-Schritt) verwenden.[75]

Die Sättigungs-Eigenschaft wird deutlich, wenn man die Formel partiell nach $tf(t,p)$ ableitet:

$$\frac{\partial BM25}{\partial tf} = IDF \times \left[\frac{k(k+1)(1 - b + b\theta)}{\left(tf + k(1 - b + b\theta)\right)^2} \right]$$

An der Gleichung ist zu erkennen, dass die Steigung von BM-25 bei größer werdenden tf (Term-Frequenz) einen abnehmenden Zuwachs aufweist, da tf nur im Nenner vorhanden ist.[76] Dies bedeutet, dass die Zunahme des BM-25 Scores mit steigender Term-Frequenz abnimmt, sodass eine Verdoppelung von 1 auf 2 einen größeren Score-Anstieg bewirkt als eine Erhöhung von 100 auf 101, was nur einer marginalen Steigerung entspricht.

Die Berücksichtigung der Passagenlänge in der Bewertung wird deutlich, wenn man die Formel partiell nach θ ableitet:

---

[72] Vgl. Cummins 2013, S. 114
[73] Vgl. Guo et al. 2020, S. 4
[74] Vgl. Robertson, Zaragoza, Taylor 2004, S. 43
[75] Vgl. Chang et al. 2020, S. 4
[76] Vgl. Robertson, Zaragoza 2009, S. 355 f.

$$\frac{\partial BM25}{\partial \theta} = IDF \times \left[\frac{-tf(k+1)kb}{\left(t + k(1-b+b\theta)\right)^2} \right]$$

An der Gleichung ist zu erkennen, dass die Steigung von BM-25 bei größer werdenden θ (das wie überdurchschnittlich lang / kurz die aktuelle Passage ist) abnimmt, da θ nur im Nenner vorkommt und der Bruch negativ ist.[77] Das heißt, dass kurze Passagen mit hoher Term-Frequenz höher bewertet werden als lange Passagen, was intuitiv ist, da längere Passagen bei gleichem Informationsgehalt mehr passende Terme enthalten sollten.[78] Das Verhalten beider Gleichungen kann sich je nach den Vorzeichen der frei wählbaren Parameter k und b ändern wobei die standard-Werte für k 1,2 und 2,0 und für b 0,75 sind.[79] Die frei wählbaren Parameter dienen den Zweck den Algorithmus für den individuellen Anwendungsfall zu optimieren.[80]

Dense Passage Retriever („Dense“ Retriever)

Der Dense Passage Retriever (DPR), ein Retrieval-System basierend auf dem Konzept der "dichten" Informationsverarbeitung, wurde im September 2020 von Karpukhin et al. vorgestellt und markiert einen signifikanten Fortschritt im Bereich des Open-Domain Question Answering (ODQA), da es signifikante Verbesserungen in der Abrufungsgenauigkeit gegenüber traditionelle sparse-retrieval-Methoden wie TF-IDF oder BM-25 aufweist.[81]

DPR adressiert das Retrieval-Problem mit der Überlegung, dass im Kern der Retrieval-Aufgabe lediglich zwei Komponenten von Bedeutung sind: die gestellte Frage und die dazugehörige Passage aus einem Dokumentenkorpus, welcher die Antwort auf ebenjene Frage enthält.[82] Hierfür nutzt DPR zwei vortrainierte BERT-Modelle als Encoder.[83] Einer dieser Encoder wird dazu verwendet, sämtliche Passagen des Korpus in dichte Vektorrepräsentationen zu encodieren und der andere Encoder transformiert die gestellte Frage in einen Vektor gleicher Dimensionalität.[84] Die beiden BERT-Encoder werden mit dem Ziel trainiert, dass die Ähnlichkeit zwischen der Frage und der Antwort-Passage im dichten Vektorraum maximiert wird.[85]

Gegeben sei ein Encoder $E_P(\cdot)$ welcher jede Passage aus dem Korpus in einen d-dimensionalen Vektor mit reellen Zahlen encodiert und ein Encoder $E_Q(\cdot)$, welcher die gestellte Frage in einen Vektor der selben Dimension encodiert. Basierend hierauf wird die Ähnlichkeit zwischen Frage q und Passage p wie folgt berechnet:[86]

[77] Vgl. Rosa et al. 2021, S. 2
[78] Vgl. Robertson, Zaragoza 2009, S. 358 ff.
[79] Vgl. Sarrouti, Ouatik El Alaoui 2017, S. 100
[80] Vgl. Robertson, Walker 1995, S. 4
[81] Vgl. Karpukhin, Oğuz, et al. 2020, S. 1
[82] Vgl. Li, Lin 2021, S. 1 ff.
[83] Vgl. Chen, Lakhotia, Oğuz, et al. 2022, S. 3
[84] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2 f.
[85] Vgl. Ren et al. 2023, S. 1 ff.
[86] Vgl. Li, Lin 2021, S. 2

$$sim(q,p) = E_Q(q)^T E_p(p)$$

Für das Maß der Ähnlichkeit wird das Skalarprodukt verwendet, wobei hier die encodierte Frage transponiert mit der encodierten Passage multipliziert wird, was zum selben Ergebnis führt.

Der gesamte Datensatz bestehend aus allen Fragen und Passagen kann in m Trainingsinstanzen unterteilt werden, wobei jede Trainingsinstanz D aus der Frage q, die richtige Passage p^+ und n irrelevante Passagen (enthält nicht die Antwort auf q) p^-, besteht:[87]

$$D = \left\{\langle q_i, {p_i}^+, {p_{i,1}}^-, \dots, {p_{i,n}}^- \rangle\right\}_{i=1}^{m}$$

Für das Training der Encoder wird eine negative log-Likelihood Funktion als Loss-Funktion verwendet, welcher zu minimieren ist:[88]

$$L\left(q_i, {p_i}^+, {p_{i,1}}^-, \dots, {p_{i,n}}^-\right) = -log\left(\frac{e^{sim(q_i, p_i^+)}}{e^{sim(q_i, p_i^+)} + \sum_{j=1}^{n} e^{sim\left(q_i, p_{i,j}^-\right)}}\right)$$

Die Loss-Funktion wird minimiert ($L = 0$), wenn die Fragen und Passagen so encodiert werden, dass es keine Ähnlichkeit zwischen der Frage und allen irrelevanten Passagen gibt ($sim(q, p^-) = 0$), da $log(1) = 0$. Aufgrund des negativen Logarithmus steigt der Loss, je größer die Ähnlichkeit zwischen Frage und irrelevanten Passagen ist, da für $\lim_{n \to 0}\left(-log(n)\right) = +\infty$.

Die Abrufung durch DPR erfolgt durch Messung der Ähnlichkeit zwischen jeder Passage und der Frage, wobei die passendsten Passagen für die Beantwortung ausgewählt werden.[89] Zusätzlich können die Passagen im Vorfeld encodiert und indexiert werden, um die Leistung des Systems bei Laufzeit zu erhöhen.[90]

Um die Genauigkeit der Abrufung insgesamt zu messen, wird die Top-K Genauigkeit des Abrufungssystems betrachtet.[91] Die Top-K-Genauigkeit misst den Anteil der an das System gestellten Fragen, bei denen die korrekte Passage unter den ersten K abgerufenen Passagen enthalten ist.[92] Experimente der Autoren haben gezeigt, dass der DPR bei der Abrufung der top 5 Passagen eine wesentlich höhere Genauigkeit aufweist als eine Abrufung mit BM-25.[93]

[87] Vgl. Karpukhin, Oğuz, et al. 2020, S. 3
[88] Vgl. Li, Lin 2021, S. 2
[89] Vgl. Chen, Lakhotia, Oğuz, et al. 2022, S. 3
[90] Vgl. Kamps et al. 2023, S. 94
[91] Vgl. Xueguang Ma et al. 2021, S. 4
[92] Vgl. Mao et al. 2021a, S. 5
[93] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2

Hybrid Retriever

Hybride Retriever sind Abrufungsmethoden welche sowohl Term-basierte sparse Algorithmen als auch Embedding-basierte dense Methoden bei der Bewertung der Passagenrelevanz zur gegebenen Frage verwenden.[94]

Die Implementierung hybrider Retriever vereint die Vorteile syntaktisch-semantischer Projektionen durch dense Embeddings mit der lexikalischen Präzision von sparse Algorithmen, wodurch die jeweiligen methodischen Schwächen ausgeglichen werden.[95] Während dense Embeddings an der Berücksichtigung spezifischer lexikalischer Details scheitern, können sparse Methoden bei lexikalischen Abweichungen, trotz semantischer Identität der Begriffe, ineffektiv sein.[96] Die Effektivität hybrider Ansätze ist Gegenstand wissenschaftlicher Debatten. Anfängliche Untersuchungen, unter anderem durch die Entwickler des Dense Passage Retrievers, suggerierten keine klare Überlegenheit hybrider Retriever im Vergleich zu reinen dense Modellen.[97] Aktuelle Forschungsergebnisse weisen jedoch auf statistisch signifikante Vorteile hybrider Modelle gegenüber reinen dense Retrievern hin, was die kontinuierliche Evolution und Evaluation in diesem Forschungsfeld unterstreicht.[98]

Gegeben sei ein Korpus K, ein Term t und eine individuelle Passage (im Korpus) p mit $p \in K$. Es ist möglich die Frage in einem binären Vektor q_v^{bm25} der Dimension $|V|$ zu erfassen, wobei V das Vokabular der Terme darstellt.[99] 1 in dem Vektor bedeutet, dass der Term Teil der Frage ist und 0, dass der Term nicht in der Frage vorkommt. Nun kann man mit dem BM-25 Algorithmus den Score aller Terme des Vokabulars in der Passage berechnen und in einen weiteren Vektor p_v^{bm25} erfassen.[100] Hierbei ist der BM-25 Score aller Terme, welche nicht in der Passage vorkommen 0. Somit ist der BM-25 Score der Passage bei gegebener Frage das Skalarprodukt der beiden Vektoren:[101]

$$BM25(Q,P) = \langle q_v^{bm25}, p_v^{bm25} \rangle$$

Wie bereits beschrieben lässt sich die Ähnlichkeit von Frage zur Passage bei dense Retrievern, z.B. im Dense Passage Retriever, ebenfalls als Skalarprodukt von (dense) Frage- und Passagen-Vektor darstellen:[102]

$$DPR(Q,P) = \langle q_v^{dpr}, p_v^{dpr} \rangle$$

---

[94] Vgl. Ji Ma et al. 2021, S. 3
[95] Vgl. Bruch, Gai, Ingber 2024, S. 4 ff.
[96] Vgl. Chen, Lakhotia, Oğuz, et al. 2022, S. 1 f.
[97] Vgl. Karpukhin, Oğuz, et al. 2020, S. 5
[98] Vgl. Ma et al. 2022, S. 614
[99] Vgl. Ji Ma et al. 2021, S. 5
[100] Vgl. Seo, Lee, Kwiatkowski, Ankur P. Parikh, et al. 2019, S. 3
[101] Vgl. Ji Ma et al. 2021, S. 5
[102] Vgl. Wang et al. 2023, S. 5

Eine Möglichkeit Hybrides Scoring zu implementieren wäre eine Linearkombination der beiden Werte wie folgt:[103]

$$sim(q^{hyb}, p^{hyb}) = \lambda\langle q_v^{bm25}, p_v^{bm25}\rangle + \langle q_v^{dpr}, p_v^{dpr}\rangle$$

λ ist ein frei wählbarer Parameter, welcher bestimmt wie viel die Lexikalische Ähnlichkeit zwischen Frage und Passage ins Gewicht der Gesamtbewertung fällt.[104]

Re-Ranking

In klassischen Retrieval-Methoden tritt das Problem auf, dass die abgerufenen Dokumente oder Passagen trotz ihrer Auswahl weiterhin irrelevante Inhalte enthalten.[105] Zudem kann die Gesamtmenge dieser Informationen so umfangreich sein, dass sie aufgrund der Performance-Anforderungen oder der begrenzten Kontextlänge des generativen Sprachmodells (LLM) für die Generierung einer Antwort als ungeeignet betrachtet wird.[106]

Eine Lösungsstrategie, die in RAG-Systemen zur Anwendung kommt, ist der Einsatz eines zusätzlichen Algorithmus, der die zuvor abgerufenen Passagen hinsichtlich ihrer Relevanz bezüglich der gestellten Frage neu ordnet (engl. re-ranking).[107] Dies ermöglicht eine effektivere Auswahl der relevantesten Passagen aus der Gesamtheit der Abrufung, wodurch die Effizienz und Präzision der Antwortgenerierung signifikant verbessert werden kann.[108]

Gemäß der vorgestellten Konzeptmatrix (siehe: Tab. 1) wird der Cross-Attention-Rescorer, auch bekannt als BERT-Rescorer, häufig für das Re-Ranking von Passagen eingesetzt.[109] Dabei werden sowohl die Anfrage als auch jede Passage in Token zerlegt und jede Anfrage mit einer Passage paarweise (mit einem Trenntoken „[SEP]“), zu einer einzigen Sequenz konkateniert.[110] Diese Sequenz wird dann einem Sprachmodell, basierend auf der Transformer-Architektur wie beispielsweise BERT, zugeführt.[111] Innerhalb des Transformer-Modells ermöglicht der Cross-Attention-Mechanismus eine wechselseitige Aufmerksamkeit zwischen jedem Token der Anfrage und jedem Token der Passage.[112] Der Zustand des speziellen "[CLS]"-Tokens, welches am Anfang der Sequenz eingefügt wird, aggregiert durch die Verarbeitung im Modell Informationen über die gesamte Sequenz.[113] Nach Durchlaufen des Modells wird die finale Einbettung des "[CLS]"-Tokens einer linearen Schicht zugeführt, die einen skalaren Wert produziert, der die Relevanz der Passage im Kontext der gestellten Anfrage

[103] Vgl. Ji Ma et al. 2021, S. 5
[104] Vgl. Chen et al. 2022, S. 5
[105] Vgl. Zhu et al. 2021, S. 4
[106] Vgl. Gao et al. 2024, S. 10 f.
[107] Vgl. Zhu et al. 2021, S. 4
[108] Vgl. Mao et al. 2021b, S. 344 ff.
[109] Vgl. Mao et al. 2021b, S. 344
[110] Vgl. MacAvaney et al. 2019, S. 2
[111] Vgl. Yang et al. 2020, S. 2
[112] Vgl. Devlin et al. 2019, S. 4 ff.
[113] Vgl. Zhan et al. 2020, S. 1 ff.

widerspiegelt.[114] Auf Basis dieses Relevanzwertes können die Passagen innerhalb der vorselektierten Menge neu geordnet werden, wobei Passagen mit höheren Relevanzwerten bevorzugt werden.[115]

Das Cross-Attention-Rescoring bietet im Kontext der Informationssuche für Fragebeantwortung in offenen und spezifischen Domänen[116] eine gesteigerte Präzision und Effektivität des Informationsabrufs.[117] Allerdings ergibt sich als problematisch, dass die Evaluierung der Relevanz von allen Passagen in Bezug auf spezifische Anfragen mit dieser Methode einen beträchtlich hohen Rechenaufwand verursacht.[118] Aus diesem Grund erweist sich das Cross-Attention-Rescoring als geeignet für das Re-Ranking von Passagen, da die Menge der Passagen durch den initialen Abrufungsprozess bereits signifikant limitiert wird.[119]

Der Rechenaufwand beim Cross-Attention-Rescoring ist intensiver, da jeder einzelne Frage Token mittels Gewichtsmatrixmultiplikationen mit jedem einzelnen Passage-Token der Passage in Beziehung gesetzt werden muss.[120] Diese Matrixmultiplikationen sind um ein Vielfaches rechenintensiver als die Prozeduren, die bei Verfahren wie DPR angewandt werden, wo die Relation von Frage und Passage lediglich am Ende durch ein Skalarprodukt der dichten Vektoren (Dense Embeddings) der Texte etabliert wird.[121] Darüber hinaus ist das Cross-Attention Scoring laufzeitbelastend, da im Gegensatz zu DPR die Passagen nicht vorab in Dense Embeddings umgewandelt und indexiert werden können.[122] Für das Cross-Attention Scoring ist die Anfrage zwingend erforderlich, was bedeutet, dass bei jeder Anfrage potenziell ein Corpus von Millionen von Dokumenten erneut durchsucht werden muss.[123]

Grundsätzlich ist es intuitiv möglich die besten Passagen (in beliebiger Anzahl) aus einer Menge von, z.B. durch DPR, bewerteten Passagen durch eine andere Methode der Relevanzbewertung neu zu ordnen. Neben der Anwendung des Cross-Attention-Rescorers für die Neuanordnung der abgerufenen Passagen, kam auch der BM-25-Algorithmus für das Re-Ranking zum Einsatz,[124] was zu unterschiedlichen Ergebnissen führte. Experimentelle Untersuchungen auf der einen Seite haben gezeigt, dass ein Re-Ranking mittels BM-25, insbesondere auf domänübergreifenden Daten nach dem initialen Abruf durch DPR, keine signifikanten

[114] Vgl. Nogueira, Cho 2020, S. 2
[115] Vgl. Nogueira, Cho 2020, S. 1 f.
[116] Vgl. Wang, Ma, Chen 2024, S. 4
[117] Vgl. Zhiguo Wang et al. 2019a, S. 4 f.
[118] Vgl. Mao et al. 2021b, S. 1
[119] Vgl. Nair et al. 2022, S. 1
[120] Vgl. Devlin et al. 2019, S. 3
[121] Vgl. Li, Gaussier 2022, S. 1
[122] Vgl. Wang, Zhuang, Zuccon 2021, S. 317 f.
[123] Vgl. Wang, Zhuang, Zuccon 2021, S. 318
[124] Vgl. Karpukhin, Oğuz, et al. 2020, S. 4 ff.

Verbesserungen hinsichtlich der Präzision der Abrufleistung erbrachte.[125] Andere Experimente haben jedoch Vorteile durch die Einbindung von BM-25 in DPR gezeigt.[126]

In der Studie von Karpukhin et al. (2020) wurde jedoch auch aufgezeigt, dass die Kombination von DPR und BM-25 in spezifischen Kontexten vorteilhaft sein kann. Dieser Vorteil ergibt sich insbesondere aus der Kapazität des BM-25-Algorithmus, relevante Terme direkt zu identifizieren und abzugleichen, eine Fähigkeit, die DPR in gewissem Maße fehlt.[127] Die Forschenden Wang et al. (2021) haben sich diesem Ansatz beträchtliche Aufmerksamkeit gewidmet und durch empirische Untersuchungen mit diversen Datensätzen demonstriert, dass ein synergetisches Zusammenspiel zwischen BERT-basierten, dichten Abrufungsmethoden und der lexikalischen Analysefähigkeit von BM-25 eine signifikante Steigerung der Abrufpräzision bewirkt, insbesondere wenn die Präzision über eine umfangreiche Menge abgerufener Passagen, beispielsweise 1000, evaluiert wird.[128] Dies lässt sich dadurch erklären, dass dichte Abrufungsmethoden effektiv in der Identifikation starker Relevanzsignale sind, jedoch Schwächen in der Erkennung subtiler Relevanzsignale aufweisen.[129] Da BM-25 eine ausgeprägte Effektivität in der Identifikation eben jener subtilen Signale zeigt, ergänzen sich die beiden Ansätze gegenseitig, indem sie die jeweiligen Limitationen ausgleichen.[130]

Daraus lässt sich die Hypothese ableiten, dass die Kombination von DPR und BM-25, eventuell auch in der Form eines BM-25 Re-Rankers, zur Bewertung der Passagenrelevanz auch bei domänenspezifischen Daten zu einer Optimierung der Abrufgenauigkeit führen könnte.

Approximate Nearest Neighbor Suche

Die Approximate Nearest Neighbor (ANN) Suche wird in der Konzeptmatrix (siehe: Tab. 1) als Methode identifiziert, die häufig zum Auffinden der relevantesten Passagen in Bezug auf eine Anfrage unter Berücksichtigung bereits berechneter Ähnlichkeitswerte eingesetzt wird.[131] Das Ziel von ANN wird wie folgt definiert:[132]

Gegeben sei eine Menge an Passagen P mit spezifischen Passagen p, eine Funktion $f_P(p)$, welcher jeder Passage p seinen entsprechenden Vektorrepräsentation zuordnet, und eine Funktion $\Psi(\phi_i, k')$ welcher die k' ähnlichsten Passagen-Vektorrepräsentationen zur Query-Vektorrepräsentation ϕ_i identifiziert. Nun ist folgende Formel zu betrachten:

$$P(\phi_i, k') = \{p \in P\colon f_P(p) \cap \Psi(\phi_i, k') \neq \emptyset\}$$

[125] Vgl. Karpukhin et al. 2020, S. 5
[126] Vgl. Glass et al. 2022, S. 9
[127] Vgl. Karpukhin, Oğuz, et al. 2020, S. 7
[128] Vgl. Wang, Zhuang, Zuccon 2021, S. 320
[129] Vgl. Wang, Zhuang, Zuccon 2021, S. 318
[130] Vgl. Wang, Zhuang, Zuccon 2021, S. 318
[131] Vgl. Xiong, Xiong, Li, Tang, Liu, Paul Bennett, et al. 2020, S. 1 f.
[132] Vgl. Macdonald, Tonellotto 2021, S. 3319

Es ist das Ziel von ANN eine Menge P zu finden, welche aus einzelnen relevanten Passagen p besteht, indem für jeden ϕ eine Anzahl von k' Vektoren gefunden wird, die basierend auf einer approximierten Distanz am nächsten zu ϕ sind (approximierte nächsten Nachbarn).[133] Mit der Zuordnungsfunktion $f_P(p)$ wird jeder der k' Vektoren ihrer ursprünglichen Passage zugeordnet.[134] Bei ANN wird allgemein nicht eine einzelne Frage betrachtet, sondern eine Menge von Fragen, da die Möglichkeit besteht, dass eine RAG-Architektur gewählt wird, wo die Frage des Anwenders synthetisch erweitert wird (engl. query-expansion).[135] Somit müssen die für jede Frage identifizierten Passagen zu einer großen Menge zusammengeführt werden um P zu erhalten:[136]

$$P(k') = \cup_{i=1}^{|q|} P(\phi_i, k')$$

In der vorliegenden Arbeit wird nur das naive RAG betrachtet weshalb davon ausgegangen werden kann, dass es für jede Suche nur eine Frage und somit nur einen ϕ gibt. Aus diesem Grund ist auch die Zusammenführung der Mengen nicht notwendig. k' ist ein wählbarer Parameter, wobei die Wahl wissenschaftlich diskutiert wird, da es nicht zwingend vorteilhaft ist, dass alle Fragen gleichermaßen zur Kandidatenauswahl der Passagen beitragen (für jede Frage kann ein anderer k Wert gesetzt werden).[137] Aufgrund des Fokus auf das naive RAG ist diese Diskussion für die vorliegende Arbeit irrelevant, k wird als einmalig frei wählbarer Parameter betrachtet.

Ein wesentliches Konzept in ANN ist die Approximation der semantischen Distanz von Frage zur Passage, anstatt die konkrete Distanz für jedes Frage-Passagenpaar auszurechnen. Intuitiv ist es möglich den Sinn der Approximation wie folgt vorzustellen:[138]

Seien P_1 und P_2 zwei dichte Passagen-Vektoren und q der dichte Frage-Vektor. Wenn im Vorhinein bekannt ist, dass P_1 und P_2 eine sich semantisch sehr ähnlich sind und q semantisch sehr unähnlich zu P_1 ist, dann kann mit relativ hoher Sicherheit davon ausgegangen werden, dass P_2 ebenfalls sehr unähnlich zu q ist, ohne die Ähnlichkeit konkret zu berechnen. Diese Approximation ermöglicht eine Abrufung mit höherer Laufzeiteffizienz als eine Direkte Suche mit Berechnung von allen Ähnlichkeitswerten von Frage zu Passagen.

Die Methodik und die genaue Ausführung der Approximation innerhalb von ANN-Algorithmen stellt ein intensiv erforschtes und breit diskutiertes Feld innerhalb der Wissenschaft dar.[139] Ein zentrales Konzept bei der Implementierung von ANN in RAG-Systemen ist die Indexierung

[133] Vgl. Malkov, Yashunin 2020, S. 1
[134] Vgl. Macdonald, Tonellotto 2021, S. 3319
[135] Vgl. Zhu et al. 2023, S. 3
[136] Vgl. Macdonald, Tonellotto 2021, S. 3319
[137] Vgl. Macdonald, Tonellotto 2021, S. 3319
[138] Vgl. Tschopp, Diggavi 2009, S. 2
[139] Vgl. Li et al. 2016, S. 1

von Passagen im Vorbereitungsschritt, wobei die Indizes so konstruiert werden, dass sie Ähnlichkeitsinformationen zwischen den Passagen enthalten, die eine Approximation der Ähnlichkeit zur Frage ermöglichen.[140] Die spezifische Methode der Approximation durch ANN beeinflusst direkt die Art und Weise der Indexierung und variiert abhängig von der gewählten Approximationsmethodik bzw. dem eingesetzten Algorithmus.[141]

2.3.3 Evaluationsmetriken

Im folgenden Kapitel werden Metriken zur Evaluation des Abrufungsergebnisses vorgestellt. Dabei wird von Metriken Abstand genommen, die eine Gesamtbewertung einschließlich anderer Faktoren wie der Antwortgenerierung durchführen und nicht explizit die reine Effektivität des Abrufungsprozesses beurteilen.

Precision

Die "Precision" Metrik P wird definiert als das Verhältnis von (abgerufenen) relevanten Passagen zu allen abgerufenen Passagen und repräsentiert die Wahrscheinlichkeit, dass eine abgerufene Passage relevant ist.[142] P wie folgt erfasst:[143]

$$P = \frac{Anzahl\ der\ Relevanten\ Passagen^{144}}{Gesamtzzahl\ der\ abgerufenen\ Passagen}$$

Diese Metrik bestraft die Abrufung von irrelevanten Passagen (falsch positiv), jedoch nicht das allgemeine Versagen des Systems Passagen abzurufen die relevant sind (falsch negativ).[145]

Average Precision

Die "Average Precision" (AP) Metrik ist eine Erweiterung der "Precision" Metrik und ist definiert als der Durchschnitt der Präzisionswerte P (innerhalb eines Abrufungsdurchlaufs für eine Frage), die nach dem Abrufen jeder relevanten Passage ermittelt werden.[146]

Mathematisch kann AP wie folgt definiert werden:[147]

$$AP = \frac{1}{R} \times \sum_{k=1}^{n} rel(k) \times P(k)$$

Dabei beschreibt R die Anzahl aller relevanten Passagen, n die Anzahl aller abgerufenen Passagen, $P(k)$ die „Precision" des Abrufungssystems an der Passage k und $rel(k)$ ist eine Indikatorfunktion die 1 ist, wenn die Passage an der stelle k relevant für die Frage ist und 0,

[140] Vgl. Chiu, Prayoonwong, Liao 2019, S. 2
[141] Vgl. Li et al. 2016, S. 3 ff.
[142] Vgl. Saracevic 1995, S. 143
[143] f. Sujatha, Dhavachelvan 2011, S. 40
[144] Aus der Menge der abgerufenen Passagen
[145] Vgl. Alvarez 2002, S. 2
[146] Vgl. Buckley, Voorhees 2017, S. 34
[147] Vgl. Cormack, Lynam 2006, S. 533

wenn die Passage irrelevant ist. Die Gesamtzahl der abgerufenen Passagen kann virtuell begrenz werden indem die Metrik nur auf die ersten K-Werte angewendet wird ($P@K$).[148] Bei einem Abrufungssystem wie DPR oder BM-25, wo die Passagen nicht nacheinander abgerufen werden, sondern alle Passagen nach Relevanz geordnet werden, beschreibt k den Rang der Passage nach der Ordnung.[149] AP wird maximiert, wenn alle relevanten Passagen des Korpus ununterbrochen die höchsten Ränge der Abrufung einnehmen.[150] Mit dieser Metrik wird die Rangordnung des Abrufungssystems bewertet, wobei irrelevante Passagen, die vor relevanten Passagen im Rang stehen und nicht erkannte relevante Passagen den AP-Wert mindern.[151] Ein Nachteil dieser Metrik besteht darin, dass die Relevanz der Passagen nur binär erfasst wird und somit die Fähigkeit eines Abrufungssystems Relevante Passagen nach ihrer Relevanz zu sortieren nicht bewertet werden kann.[152]

Mean Average Precision

Um AP über eine Menge von bearbeiteten Anfragen zu berechnen kann die Mean Average Precision (MAP) Metrik verwendet werden.[153] MAP wird berechnet, indem der Mittelwert von AP über N Fragen berechnet wird:[154]

$$MAP = \frac{1}{N} \times \sum_{j=1}^{N} \left(\frac{1}{R} \times \sum_{k=1}^{n} rel(k) \times P(k) \right)$$

Recall

Im gegensatz zur “Precision” misst Recall (R) den Anteil der relevanten Passagen, die tatsächlich aubgerufen wurden, im Verhältnis zu allen existierenden relevanten Passagen im Korpus.[155]

Recall wird wie folgt berechnet: [156]

$$R = \frac{Anzahl\ relevanter\ Passagen\ die\ abgerufen\ wurden}{Anzahl\ aller\ relevanten\ Passagen\ im\ Korpus}$$

Um zu vermeiden, dass immer Recall = 1 gilt, da bei der Abrufung aller Passagen alle relevanten passagen zwingend enthalten ist, wird oft Recall bei einer festen Anzahl von abgerufenen Passagen (oder ersten n Ränge) wie z.B. Recall@1000,

[148] Vgl. Sujatha, Dhavachelvan 2011, S. 40 f.
[149] Vgl. Xu, Li 2007, S. 393
[150] Intuitiv verständlich bei der Durchführung einer beispielhaften Berechnung
[151] Vgl. Kishida 2005, S. 4
[152] Vgl. Kishida 2005, S. 2
[153] Vgl. Sujatha, Dhavachelvan 2011, S. 40
[154] Vgl. Korra et al. 2011, S. 724
[155] Vgl. Buckley, Voorhees 2017, S. 34
[156] Vgl. Arora, Kanjilal, Varshney 2016, S. 227

gemessen.[157] Diese Metrik bestraft das System bei allgemeinen Nicht-Abrufen von relevanten Passagen (falsch negativ).[158]

F1

Die F1-Metrik, welcher eine Kombination aus “Precision” und Recall darstellt, wird als das harmonische Mittel dieser beiden Metriken definiert.[159] Mathematisch wird F1 wie folgt erfasst:[160]

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Um eine Division durch 0 zu vermeiden, wenn das Abrufungssystem keine einzige relevante Passage abruft, kann die Formel wie folgt umgeschrieben werden:[161]

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

Auch bei F1 bietet es sich an die Metrik auf eine bestimmte Menge von abgerufenen Passagen anzuwenden wie z.B. F1@100.[162]

Accuracy (Top-K)

Die „accuracy“-Metrik (hier A) misst die Genauigkeit eines Abrufungssystems, indem sie den Anteil jener Abrufdurchläufe auf einer Menge von Fragen ermittelt, bei denen unter den abgerufenen Passagen mindestens eine Passage die Antwort auf die Frage beinhaltet.[163] Somit kann sie Mathematisch wie folgt beschrieben werden:[164]

$$A = \frac{Anzahl\ der\ durch\ die\ Abrufung\ beantworteten\ Fragen}{Gesamtzahl\ der\ Fragen}$$

Um die Bewertung des Systems spezifischer zu gestalten, wird auch bei dieser Metrik oft ein Grenzwert K eingeführt, wo die Abrufung nur bis zu den Top K Passagen betrachtet wird und die beantwortende Passage sich darunter befinden muss.[165] Diese Variation der Metrik wird als „Top-k accuracy“ bezeichnet.[166] Diese Metrik wird besonders bei Datensätzen angewandt, bei denen im Vorhinein konkrete Passagen definiert werden können, die als sogenannte „Gold“-Passagen dienen und die Fragen beantworten, auf welche getestet wird.[167]

[157] Vgl. Buckley, Voorhees 2017, S. 34
[158] Vgl. Alvarez 2002, S. 2
[159] Vgl. McSherry, Najork 2008, S. 418
[160] Vgl. Yedidia 2016, S. 1
[161] Vgl. McSherry, Najork 2008, S. 418
[162] Vgl. McSherry, Najork 2008, S. 418
[163] Vgl. Mallen et al. 2023, S. 3
[164] Vgl. Soudani, Kanoulas, Hasibi 2024, S. 3 ff.
[165] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2
[166] Vgl. Lee, Wettig, Chen 2021a, S. 3
[167] Vgl. Karpukhin, Oğuz, et al. 2020, S. 6 f.

Exact Match

Die Exact Match (EM) Metrik misst klassisch die genaue lexikalische Übereinstimmung der von RAG-Systemen generierten Antwort auf der Basis abgerufener Passagen mit gegebenen Musterantworten.[168] Obwohl EM dazu verwendet werden könnte, zu messen, wie oft ein Abrufsystem bei einer gestellten Frage Passagen ermitteln kann, die lexikalisch genau die richtige Antwort entsprechend der Musterantwort liefern,[169] wird dieser Ansatz in der vorliegenden Arbeit nicht näher betrachtet, da in der identifizierten relevanten Forschungsliteratur kein solcher Fall beschrieben wurde.

Mean Reciprocal Rank

Die Mean Reciprocal Rank (MRR) Metrik bewertet ein Abrufungssystem basierend auf der Position der ersten relevanten Passage in der Rangordnung der Relevanz zur Frage (RR), gemittelt nach einer Menge von Fragen.[170] Die MRR-Metrik wird mathematisch wie folgt definiert:[171]

$$MRR = \frac{1}{Q} \times \sum_{i=1}^{Q} \frac{1}{RR_i}$$

Dabei beschreibt Q die Menge aller Fragen und RR_i den Rang der ersten relevanten Passage wobei MRR maximiert wird, wenn auf jeder Frage die höchst bewertete Passage auch tatsächlich relevant ist.[172] Diese Metrik ist besonders relevant bei der Bewertung von Abrufungen in Systemen, wo die Menge der betrachteten Passagen niedrig ist (z.B. Top-k, k = 3).[173]

Normalised Discounted Cumulative Gain

Eine weitere Metrik zur Evaluation der Relevanzbewertung und -ordnung von Abrufungssystemen stellt der Normalized Discounted Cumulative Gain (NDCG) dar, der sich auf eine definierte Menge der Top-K abgerufenen beziehungsweise nach Relevanz geordneten Passagen anwendet.[174] Mathematisch lässt sich NDCG wie folgt berechnen:[175]

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

$$DCG@K = \sum_{i=1}^{k} \frac{2r_i - 1}{log(i+1)}$$

[168] Vgl. Mao et al. 2021a, S. 5
[169] Vgl. Chen et al. 2017a, S. 6
[170] Vgl. Shi et al. 2012, S. 140
[171] Vgl. Chavhan, Raghuwanshi, Dharmik 2021, S. 5
[172] Vgl. Chavhan, Raghuwanshi, Dharmik 2021, S. 5
[173] Vgl. Shi et al. 2012, S. 140
[174] Vgl. Valizadegan et al. 2009, S. 1 f.
[175] Vgl. Chavhan, Raghuwanshi, Dharmik 2021, S. 5

Dabei repräsentiert k die Menge der betrachteten Dokumente, r_i der bestimmte Relevanzwert der Passage an der Position i und $IDCG$ der DCG-Wert der Passagenmenge k, wenn die Passagen in der richtigen Relevanzreihenfolge stehen würden.[176] An der Formel ist zu erkennen, dass $NDCG$ maximiiert wird, wenn das Abrufungsystem Passagen in der genau richtigen Relevanzreihenfolge abruft, wobei der Maximale $NDCG$-Wert aufgrund des Normalisierungsfaktors $IDCG$ 1 beträgt ($DCG = IDCG$).[177]

2.3.4 Synthetische Generierung von Daten

In der Konzeptmatrix (siehe: Tab. 1) wurden verschiedene Konzepte im Bereich der synthetischen Generation von Daten identifiziert. Zum einen die synthetische Anreichreicherung der Frage mit von Sprachmodellen generierten Inhalten[178] und die synthetische Generierung von Trainings- und Testdaten, insbesondere die synthetische Generierung von Fragen.[179] Aufgrund des Fokus der vorliegenden Arbeit auf das naive RAG wird die synthetische Anreicherung der Frage nicht weiter betrachtet.

Synthetische Generierung von Fragen

Ein Hindernis bei der Entwicklung und Evaluation von Abrufungssystemen im Zero-Shot Szenario,[180] insbesondere mit dem Ziel, auf domänenspezifische Anfragen präzise Textpassagen zu ermitteln oder mit domänenspezifischen Inhalten zu interagieren, besteht in der mangelnden Verfügbarkeit adäquater Frage-Antwort-Paare.[181] Hierbei wird unter einer Antwort der Kontext oder die Textpassage verstanden, die die gestellte Frage inhaltlich treffend beantwortet.[182] Diese Herausforderung erweist sich nicht nur akademisch als relevant, sondern auch in der praktischen Anwendung von RAG-Systemen in unternehmensspezifischen Kontexten, wo das primäre Ziel darin bestehen kann, unternehmensinterne Daten oder Dokumente zur Beantwortung von Fragen zu nutzen.[183] Eine verbreitete Methodik zum Training oder zur Evaluation dieser Systeme involviert die Nutzung öffentlich zugänglicher Daten- oder Benchmarking-Sätze, welche Fragen, den jeweiligen Kontext bzw. die antwortende Passage umfassen.[184] Dennoch repräsentiert auch dieser Ansatz keine ideale Lösung, da einerseits die verfügbaren Datensätze oftmals nicht die spezifische Domäne des Einsatzbereichs des RAG-Systems abdecken[185] und andererseits in der Forschung die Auffassung vertreten wird, dass insbesondere bei auf dichten Vektorrepräsentationen basierenden Abrufungssystemen die

[176] Vgl. Radlinski, Craswell 2010, S. 669
[177] Vgl. Valizadegan et al., S. 2
[178] Vgl. Wang, Ma, Chen 2024, S. 3
[179] Vgl. Alberti et al. 2019, S. 1
[180] Vgl. Xian et al. 2020, S. 1 f.
[181] Vgl. Ji Ma et al. 2021b, S. 1 f.
[182] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2
[183] Vgl. Cleverley, Burnett 2019, S. 64 f.
[184] Vgl. Lewis et al. 2020, S. 4
[185] Vgl. Li et al. 2022a, S. 2

Leistungsfähigkeit hinsichtlich eines spezifischen Datensatzes nicht zwangsläufig auf andere, bisher unberücksichtigte Datensätze übertragbar ist.[186]

Ein Ansatz zur Bewältigung dieser Herausforderung besteht in der synthetischen Generierung von Fragen.[187] Dieser Prozess involviert das Einpflegen von Textpassagen, die den zugrundeliegenden Informationskorpus konstituieren, in ein Sprachmodell, welches dazu angeleitet wird, eine inhaltlich passende Frage zu formulieren, die von der Passage beantwortet werden kann.[188] Durch dieses Verfahren ist es möglich, qualitativ hochwertige Datensätze zu erstellen, die auf Dokumenten oder Informationen basieren, welche in der Produktivumgebung des Abrufungssystems verwendet werden sollen, oder repräsentativ dafür sind.[189] Diese Datensätze können dann für die Evaluierung des Systems oder Training einzelner Komponente verwendet werden.[190] Während es möglich ist, das für die Fragegenerierung zuständige Sprachmodell mittels vorhandener Frage-Passage-Paare aus öffentlichen Datensätzen zu trainieren,[191] geht die vorliegende Arbeit von der Verwendung bereits vortrainierter und zugänglicher Sprachmodelle im Kontext der synthetischen Fragegenerierung aus.[192]

In der nachfolgenden Abbildung (siehe: Abb. 2) sind zwei Beispiele dargestellt, bei denen basierend auf den Passagen in einem Zero-Shot-Szenario passende Fragen generiert werden. Bei der ersten Frage wird ersichtlich, dass eine Antwort direkt aus dem Inhalt der Passage abgeleitet werden kann, wobei als mögliche Antwort formuliert werden könnte: „Walgreens plans to replicate its zero waste to landfill program throughout its network of 17 distribution centers in the USA and one in Puerto Rico.“ Passage 2 präsentiert eine Tabelle, die aus einem PDF-Dokument extrahiert wurde, in dem die Tabelle ursprünglich enthalten war. Diese Extraktion erfolgte mittels der TableFormer-Methodik, einem Transformer-basierten Verfahren, das darauf abzielt, Tabellen und andere strukturierte Textelemente direkt aus PDF-Dokumenten zu lesen.[193] Als ergänzendes Konzept existiert die Nutzung rekurrenter neuronaler Netze zur Umwandlung von PDF-Dokumenten in durchsuchbaren Text, was die Gewinnung nutzbarer Textpassagen aus unternehmens- oder organisationsinternen Dokumenten erleichtert.[194] Die im Kontext der zweiten Passage generierte Frage kann eindeutig durch die Informationen in der Tabelle beantwortet werden. Eine mögliche Antwort hierauf könnte lauten: „The disclosure title that corresponds to the page numbers 7-21 in the 2021 10-K report is 'Conflicts of interest'."

[186] Vgl. Khramtsova et al. 2023a, S. 1
[187] Vgl. Ji Ma et al. 2021, S. 1
[188] Vgl. Zhou et al. 2018, S. 1
[189] Vgl. Ji Ma et al. 2021, S. 1
[190] Vgl. Oğuz et al. 2021, S. 2 ff.
[191] Vgl. Zhou et al. 2018, S. 2 f.
[192] Vgl. Sachan et al. 2023, S. 600
[193] Vgl. Nassar et al. 2022a, S. 8
[194] Vgl. Livathinos et al. 2021, S. 15137 f.

Passage 1:
The Walgreens distribution center in Moreno Valley, California, USA, launched a zero waste to landfill pilot program in fis cal 2016. The pilot aims to contribute directly to our CSR environmental sustainability initiatives. Once fully developed into an effective multi-stream recycling and waste reduction program, Walgreens aims to replicate the program throughout its network of 17 distribution centers in the USA and one in Puerto Rico.

Generierte Frage 1:
Where does Walgreens plan to replicate its zero waste to landfill program once it's fully developed?

Passage 2:
|----------|----------|----------|----------|
GRI Disclosure Number	Disclosure Title	Page	Reference/Location
102-24	Nominating and selecting the highest governance body	14,15	2021 Schedule 14-A
102-25	Conflicts of interest	7-21	2021 10-K
102-26	Role of highest governance body in setting purpose, values, and strategy	14, 15	2021 Schedule 14-A
102-29	Identifying and managing economic, environmental, and social impacts	2-18	2021 ESG Report
102-45	Entities included in the consolidated financial statements	1	2021 ESG Report
102-50	Reporting period	1, 2	2021 10-K
102-53	Contact point for questions regarding the report	22	2021 ESG Report
201-1	Direct economic value generated and distributed	8-11	2021 Annual report
201-2	Financial implications and other risk and opportunities due to climate change	11	2021 ESG Report
401-2	Benefits provided to full-time employees that are not provided to temporary or part- time employees	5 , 13	2021 10-K & 2021 ESG Report
413-1	Operations with local community engagement, impact, assessments and development programs	14	2021 ESG Report

Generierte Frage 2:
Which disclosure title corresponds to the page numbers 7-21 in the 2021 10-K report?

Abb. 2: Beispiele generierter Fragen aus unternehmenseigener Implementierung

2.3.5 Diskussion der Konzepte im Hinblick auf die praktische Umsetzung

In der aktuellen Forschungslandschaft stellen Abrufungssysteme, die auf dichten Vektorrepräsentationen basieren, die bevorzugte Methode für die Informationsabrufung dar.[195] Diese Dominanz ist ihrer überlegenen Fähigkeit zuzuschreiben, semantische Ähnlichkeiten zwischen Fragen und Textpassagen präziser zu erfassen als Systeme, die sich ausschließlich auf lexikalische Analysen („sparse“ Abrufungsmethoden) stützen.[196] Obwohl das DPR-System, entwickelt von Karpukhin et al., den Kern vieler moderner RAG-Systeme bildet (siehe: Tab. 1), war es nicht das erste System, das eine höhere Leistung als lexikalische bzw. spärliche Abrufungsmethoden aufwies.[197] Diese Leistung wurde erstmals vom „OpenRetrieval Question Answering System“ (ORQA) von Lee et al., vorgestellt im Jahr 2019, erreicht.[198] Durch diese

[195] Vgl. Tong Chen et al. 2023, S. 1
[196] Vgl. Fu et al. 2023, S. 1 f.
[197] Vgl. Karpukhin, Oğuz, et al. 2020, S. 1
[198] Vgl. Lee, Chang, Toutanova 2019a, S. 1

Innovation konnten die Forschenden erstmals die zuvor dominierende Methode BM-25 in Bezug auf die Effektivität der Informationsabrufung übertreffen.[199] Jedoch findet ORQA derzeit keine breite Anwendung, was hauptsächlich auf die Notwendigkeit eines aufwendigen vorgeschalteten Trainings mittels der Inverse-Cloze-Aufgabe zurückzuführen ist.[200] Im Gegensatz dazu erfordert DPR kein solch kostspieliges Vorab-Training und erreicht dennoch eine signifikante Überlegenheit gegenüber BM25-basierten Systemen.[201] In der Forschung wird die Kostenüberlegung ebenfalls in die Abwägung einbezogen, ob der Aufwand für das Vorabtraining dichter Abrufungssysteme im Vergleich zu spärlichen Systemen gerechtfertigt ist.[202] Die zuvor diskutierte Praxis der hybriden Abrufungssysteme repräsentiert einen Ansatz, um das Gleichgewicht zwischen den Kosten und Nutzen dichter und spärlicher Abrufungssysteme zu optimieren.[203] Wie bereits in Kapitel 2.3.2 erörtert, ist die Integration von spärlichen und dichten Abrufmethoden nicht nur aus Kostengründen relevant, sondern trägt auch signifikant zur Verbesserung der Abrufungsergebnisse bei. In der Praxis lässt sich der Kostenaufwand für das Vorabtraining durch den Einsatz öffentlich zugänglicher und vortrainierter Embedding-Modelle reduzieren, wie zum Beispiel das BAAI/bge-small-en-v1.5 Modell verfügbar auf der Platform "Hugging Face".[204] Allerdings besteht weiterhin die Herausforderung, dass die Genauigkeit und Leistungsfähigkeit dieser Modelle bei der Erfassung semantischer Bedeutungen von Texten nicht zwangsläufig von den ursprünglichen Trainingsdatensätzen auf die Daten in einem spezifischen Anwendungskontext übertragbar sind.[205] Dennoch ist dieser Ansatz von Bedeutung, da es in der Praxis oft erforderlich ist, generalisierbare und universell einsetzbare Lösungen zu entwickeln, statt auf Einzelfälle zugeschnittene Anwendungen zu erstellen.[206] Dies bedeutet, dass ein spezifisches Vorabtraining auf den produktiven Datensatz aus systemtechnischer und Produkt Perspektive nicht immer möglich ist. Zudem kann es Situationen geben, in denen nicht ausreichend Ressourcen für ein umfassendes Vorabtraining oder Feinabstimmung dieser Modelle zur Verfügung stehen. Die zuvor diskutierte ANN-Suche leistet auch in der Praxis einen wesentlichen Beitrag zur Verbesserung der Effizienz von Abrufsystemen.[207] Anstelle der manuellen Implementierung eines ANN-Algorithmus bieten Plattformen wie Elasticsearch Zugang zu Softwarebibliotheken, die bereits implementierte ANN-Algorithmen zur direkten Nutzung enthalten.[208] Ergänzend dazu bietet Elasticsearch

[199] Vgl. Lee, Chang, Toutanova 2019, S. 8
[200] Vgl. Widodo 2023, S. 337 f.
[201] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2
[202] Vgl. Arabzadeh, Yan, Clarke 2021, S. 1 f.
[203] Vgl. Arabzadeh, Yan, Clarke 2021, S. 1
[204] Vgl. Alquaary, Çelebi 2023, S. 81 ff.
[205] Vgl. Khramtsova et al. 2023, S. 1
[206] Vgl. Du et al. 2020a, S. 1
[207] Vgl. Xiong, Xiong, Li, Tang, Liu, Paul Bennett, et al. 2020, S. 1
[208] Vgl. Carrara et al. 2022, S. 1

auch Funktionen für eine Relevanzbewertung mittels BM-25[209] und weitere Suchfunktionen an, welche als Re-Ranker nach einer initialen Abrufung verwendet werden können.[210]

Die Wahl der Metriken für die Evaluation des Abrufungssystems sollte ebenfalls der spezifischen Ausrichtung der praktischen Implementierung entsprechen.[211] Ausgehend vom Kontext der synthetischen Fragegenerierung, bei der zu jeder Passage eine Frage erstellt wird, gilt die Annahme, dass es zu jeder Frage genau eine relevante Passage gibt.[212] In dieser Konstellation erweisen sich die "precision"-basierten Metriken (Precision, AP, MAP) als bedingt geeignet, denn bei nur einer relevanten Passage je Frage fällt der Precision-Wert binär aus.[213] Folglich ist es sinnvoll, den Precision-Wert (aggregierter Durchschnitt über alle Fragen) für eine definierte Anzahl der top-abgerufenen Passagen zu ermitteln, beispielsweise Precision@10[214] oder 1.[215] Dies verdeutlicht, in wie vielen Fällen das System in der Lage ist, die korrekte Passage konsequent an die erste Stelle zu setzen. Darauf aufbauend lässt sich feststellen, dass die AP im Kontext einer binären Auswertung stark von der Position abhängt, an welcher die relevante Passage innerhalb einer nach Relevanz sortierten Liste erscheint.[216] Dadruch wäre MAP als Bewertungsmetrik unter den gegebenen Umständen ebenfalls geeignet. Weiterhin ist der Recall-Wert, ähnlich dem Precision-Wert, in einem Szenario mit genau einer relevanten Passage pro Frage als binär anzusehen. Da in solch einem Fall Recall und Precision identische Werte annehmen würden, verliert der F1-Wert, der das harmonische Mittel von Recall und Precision darstellt, an Aussagekraft.[217] Die Metriken MRR und NDCG würden ebenfalls die Position der relevanten Passage bewerten und ihren größten Wert erreichen, wenn diese an erster Stelle steht. Jedoch werden beide Metriken (auch ungemittelt) in skalare statt Binäre Werte ausgedrückt.

Im vorliegenden spezifischen Kontext wird in der Wissenschaft auch die Top-K Accuracy als Metrik eingesetzt.[218] Diese Metrik quantifiziert, wie viele Fragen innerhalb einer definierten Anzahl von Top-K betrachteten Passagen durch das Abrufungssystem korrekt beantwortet werden können.[219] Anders als die oben genannten Metriken informiert die Top-K-Genauigkeit nicht über die genaue Position der relevanten Passage innerhalb der betrachteten Anzahl von Passagen, sondern bestätigt lediglich deren Präsenz.[220] Diese Erörterung der Evaluierungs-

[209] Vgl. Yang, David D Lewis, et al. 2018, S. 1
[210] Vgl. Kathare, O. Vinati, Prabhu 2022, S. 35 f.
[211] Vgl. Ijesunor Akhigbe, Samuel Afolabi, Rotimi Adagunodo 2011, S. 6
[212] Vgl. Ji Ma et al. 2021, S. 4
[213] Intuitiv verständlich bei beispielhafter Berechnung.
[214] Vgl. Sujatha, Dhavachelvan 2011, S. 41
[215] Vgl. Ji Ma et al. 2021, S. 8
[216] Siehe Formel in Kapitel 2.3.2
[217] Deutlich bei beispielhafter Berechnung
[218] Vgl. Karpukhin, Oğuz, et al. 2020, S. 5
[219] Vgl. Lee, Wettig, Chen 2021a, S. 3 f.
[220] Vgl. Karpukhin, Oğuz, et al. 2020, S. 2

methoden führt zur Debatte über die Relevanz der Platzierung der korrekten Passage im Hinblick auf die Gesamteffektivität eines RAG-Systems.[221] Während Metriken, welche die Position der korrekten Passage berücksichtigen, die Präzision des Abrufungssystems effektiv messen, mag deren Bedeutung für das Gesamtergebnis des RAG-Systems begrenzt sein. Dies liegt daran, dass das nachfolgende Sprachmodell, welches für die Generierung der Antwort verantwortlich ist, eine beträchtliche Kontextlänge berücksichtigen kann.[222] In der Praxis ist es möglich dem Sprachmodell die maximale Anzahl an Passagen zuführen, die zusammen mit der Frage verarbeitet werden kann.[223] Theoretisch wäre somit die exakte Position einer Passage innerhalb der Top-K Ergebnisse weniger entscheidend, sofern sie sich unter den Passagen befindet, die vom Sprachmodell aufgenommen werden können und zur Beantwortung der Frage beitragen.[224] Es gibt jedoch auch Forschende, welche die Position vertreten, dass Sprachmodelle Informationen besser verarbeiten, wenn diese sich am Anfang oder Ende des eingegebenen Kontextes befinden.[225] Dies spricht für die Relevanz von Rang/Positionsbasierten Evaluationsmetriken für das gesamtheitliche RAG-System und demnach auch für die Abrufung.

2.4 Semantische Klassifikation von Fragen

Vor der Einführung von RAG-Systemen war die Klassifizierung von Fragen nach semantischen Kriterien ein wesentlicher Bestandteil klassischer Frage-Antwort-Systeme.[226] In einem solchen System, das sich aus den Phasen der Frage-, Dokumenten- und Antwortverarbeitung zusammensetzt, erfolgt zunächst die Klassifizierung der eingegebenen Frage im Rahmen der Frageverarbeitung (siehe: Abb. 3).

Diese Klassifizierung wird allerdings nicht bei der Bewertung der Relevanz oder dem Abruf der Dokumente berücksichtigt, stattdessen werden traditionelle Methoden der Relevanzbewertung wie beispielsweise TF-IDF oder BM-25 eingesetzt.[227] Ziel dieser Methodik ist es, auf Basis der Frageklassen, zum Beispiel durch regelbasierte Zuordnungen oder Klassifikationen, die entsprechenden Antworttypen zu bestimmen.[228] Diese Antworttypen dienen dann dazu, im Rahmen der Antwortverarbeitung (Antwortidentifikation), aus einem Pool vorselektierter und bezüglich ihrer Relevanz bewerteter Dokumente die Textbestandteile zu extrahieren,[229] die potenziell zur Beantwortung der Frage beitragen könnten.[230]

---

[221] Vgl. Wu et al. 2019, S. 1
[222] Vgl. Li et al. 2023, S. 2 ff.
[223] Vgl. Lewis et al. 2020, S. 2 ff.
[224] Vgl. Gao et al. 2024, S. 4
[225] Vgl. Li et al. 2023, S. 7
[226] Vgl. Mishra, Jain 2016, S. 347 f.
[227] Vgl. Sarrouti, Ouatik El Alaoui 2017, S. 97 ff.
[228] Vgl. Kolomiyets, Moens 2011, S. 4
[229] Vgl. Ali Mohamed Nabil Allam, Haggag 2012, S. 3 ff.
[230] Vgl. Sarrouti, Ouatik El Alaoui 2017, S. 98

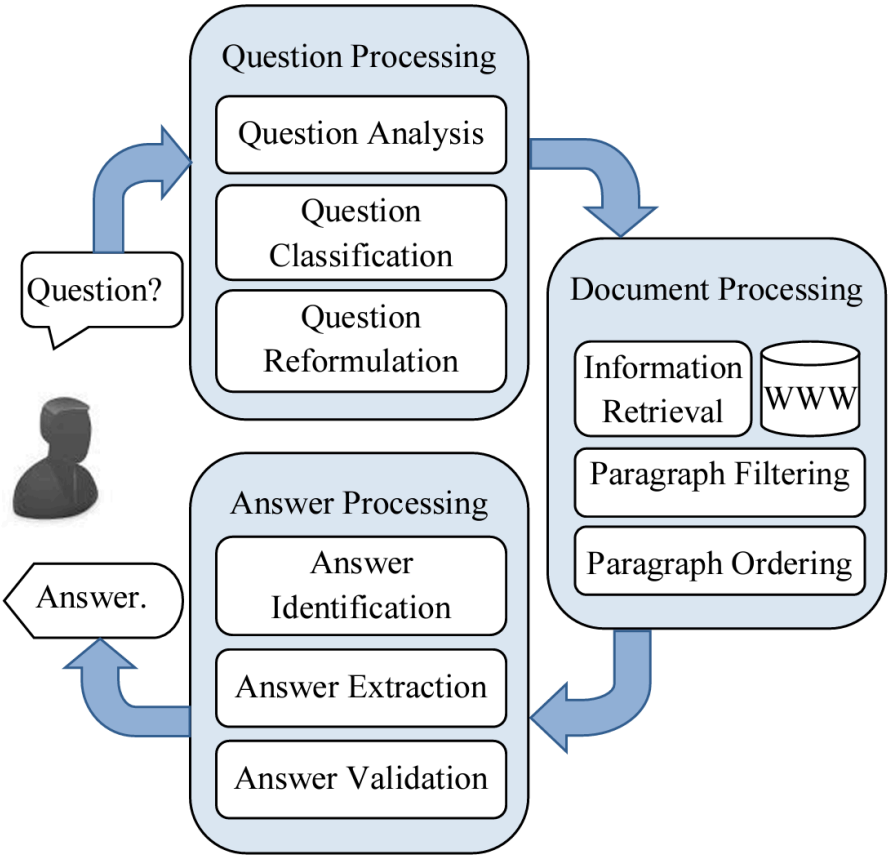


Abb. 3: Schematische Abbildung eines klassischen Frage-Antwort-Systems[231]

In der semantischen Klassifizierung von Fragen kommen unterschiedliche Taxonomien zum Einsatz.[232] Eine intuitive Methode ist die Einteilung in W-Fragen wie „was“, „wann“, „wer“, „wie“, „warum“, „wo“ und „welches“, die sich durch einfaches Suchen nach Schlüsselwörtern im Satz ohne komplexe Klassifikationsalgorithmen realisieren lässt.[233] Es existieren zudem speziell für Frage-Antwort-Systeme entwickelte flache Taxonomien, wie die von Radev et al. vorgeschlagene, basierend auf 17 semantischen Klassen.[234] Zudem wurden in der Forschung hierarchische Taxonomien empfohlen, beispielsweise die von Moldovan et al., welcher die W-Fragen als Ausgangspunkt der Klassifizeirung verwendet.[235] Diese weist jeder Frage-Klasse ebenfalls spezifischen Antwort-Klassen zu.[236] Ein bedeutender Akteur in diesem Forschungsfeld ist die „Text Retrieval Conference“ (TREC), initiiert im Jahr 1992 und gesponsort durch die „Defense Advanced Research Projects Agency“ (DARPA) und das „National Institute of Standards and Technology“ (NIST).[237] Ziel dieser Konferenz ist es, Forschung im Bereich des Informationsabrufs auf großen Datensammlungen zu fördern und

[231] Ali Mohamed Nabil Allam, Haggag 2012, S. 3
[232] Vgl. Allam, Haggag 2012, S. 5 f.
[233] Vgl. Mishra, Jain 2016, S. 352 ff.
[234] Vgl. Radev et al., S. 412
[235] Vgl. Dan Moldovan et al. 1999, S. 3
[236] Vgl. Ali Mohamed Nabil Allam, Haggag 2012, S. 6
[237] Vgl. Harman 1993, S. 36

eine Evaluationsinfrastruktur zu schaffen.[238] Sie bietet zudem eine Plattform für Forschende, um neueste Erkenntnisse und Ergebnisse zu präsentieren.[239] Seit ihrem Beginn mit 25 teilnehmenden Gruppen[240] bis zu 73 Gruppen aus 27 Ländern im Jahr 2021[241] hat die TREC-Konferenz kontinuierlich an Bedeutung und Einfluss gewonnen. Zusammen mit der anhaltenden Verwendung von TREC-Datensätzen[242] verdeutlicht dies ihre aktuelle Relevanz in der wissenschaftlichen Gemeinschaft.

Die folgende Arbeit betrachtet folgende Taxonomie von Li & Roth (2002), welches in der 10. Iteration der TREC-Konferez (TREC-10) vorgestellt wurde:

ABBREVIATION	Letter	Description	**NUMERIC**
Abbreviation	Other	Manner	Code
Expression	Plant	Reason	Count
ENTITY	Product	**HUMAN**	Date
Animal	Religion	Group	Distance
Body	Sport	Individual	Money
Color	Substance	Title	Order
Creative	Symbol	Description	Other
Currency	Technique	**LOCATION**	Period
disease medicine	Term	City	Percent
Event	Vehicle	Country	Size
Food	Word	Mountain	Speed
Instrumend	**DESCRIPTION**	Other	Temp
Language	Definition	State	Weight

Tab. 2: Hierarchische Taxonomie für semantische Klassen[243]

Die vorgestellte Taxonomie ist hierarchisch strukturiert.[244] Die fett hervorgehobenen Klassen repräsentieren die sechs übergeordneten Kategorien, während die übrigen die fünfzig

[238] Vgl. Soboroff 2022, S. 1
[239] Vgl. Craswell et al. 2020, S. 4 ff.
[240] Vgl. Harman 1993, S. 36
[241] Vgl. Soboroff 2022, S. 1
[242] Vgl. Otero, Parapar, Barreiro 2023, S. 3
[243] Vgl. Li, Roth 2002, S. 3
[244] Vgl. Sangodiah, Muniandy, Heng 2005, S. 388

detaillierten Klassen darstellen.[245] Eine Frage kann daher zwei Labels erhalten: eines für die übergeordnete und eines für die detaillierte Klasse.[246] Die Klassen reflektieren die Art der Antwort oder der relevanten Informationen, die zu der jeweiligen Klasse gehören.[247] Bei genauerer Betrachtung der Klassen wird deutlich, dass einige Fragen mehrdeutig sein können und zwei verschiedenen Klassen zugeordnet werden könnten. Beispielsweise könnte die Frage „What is the Kelvin scale?“ sowohl eine Beschreibung als auch einen numerischen Wert erfordern. Die Forschenden schlagen als Lösung für dieses Problem vor, Fragen mehreren Klassen zuzuweisen.[248]

Diese Taxonomie ist effektiv, da die Klassen domänenübergreifend und nicht zu themenspezifisch sind,[249] insbesondere die übergeordneten Klassen, die theoretisch auf alle Arten von Fragen anwendbar sind. Darüber hinaus wurde diese Taxonomie mit Fragen aus TREC-10 getestet, wobei eine Klassifikationsgenauigkeit von 95 bis 98.8 % erreicht wurde.[250] Dies belegt, dass die Taxonomie gut gewählt ist, denn sie ermöglicht es dem Lern- und Klassifikationsmodell, die Fragen mit hoher Präzision zu kategorisieren. Zudem sind qualitativ hochwertige Daten in großer Menge als öffentlich zugängliche, vorklassifizierte Fragen verfügbar, die zum Trainieren und Testen genutzt werden können.[251] Diese Vorteile tragen dazu bei, dass diese einer der am häufigsten eingesetzten Taxonomien ist.[252]

Ein weiterer relevanter Aspekt ist der "General Language Understanding Evaluation"-Benchmark (GLUE), der 2019 von Wang et al. etabliert wurde. Der GLUE-Benchmark ist eine umfassende Sammlung von Aufgaben aus dem Bereich der natürlichen Sprachverarbeitung, die dazu dient, Systeme auf ihre Fähigkeit zu testen, natürliche Sprache zu verstehen.[253] Besonders relevant für das Anwendungsfeld von Frage-Antwort-Systemen ist die Aufgabe der “Lexical Entailment“, die auf dem "Question-answering Natural Language Inference" (QNLI)-Datensatz des GLUE-Benchmarks basiert.[254] Dieser umfasst eine Vielzahl von Datenpunkten, bestehend aus einer Frage und einem zugehörigen Satz.[255] Die Aufgabe umfasst die Evaluierung, ob ein Satz die Antwort auf die zugehörige Frage beinhaltet, wobei das Label "entailment" zugewiesen wird, sofern eine Übereinstimmung festgestellt wird, und das Label "not_entailment", sofern keine Übereinstimmung vorliegt.[256]

---

[245] Vgl. Ali Mohamed Nabil Allam, Haggag 2012, S. 6
[246] Eine beispielhafte Implementierung dieser 2 Klassen: https://huggingface.co/datasets/trec
[247] Vgl. Li, Roth 2002, S. 1
[248] Vgl. Li, Roth 2002, S. 3
[249] Vgl. Madabushi, Lee, Barnden, S. 3284
[250] Vgl. Li, Roth 2002, S. 5
[251] Vgl. Xu et al. 2019, S. 1 f.
[252] Vgl. Madabushi, Lee, Barnden, S. 3284
[253] Vgl. Wang et al. 2019, S. 1
[254] Vgl. Wang et al. 2019, S. 4
[255] Vgl. Wang et al. 2022, S. 5
[256] Vgl. Alex Wang et al. 2019, S. 4 ff.

Frage	Satz	Label
What is the name of the village 9 miles north of Calafat where the Ottoman forces attacked the Russians?	On 31 December 1853, the Ottoman forces at Calafat moved against the Russian force at Chetatea or Cetate, a small village nine miles north of Calafat, and engaged them on 6 January 1854.	entailment
What was the name of the airport the United States built on Ascension Island?	A local industry manufacturing fibre from New Zealand flax was successfully reestablished in 1907 and generated considerable income during the First World War.	not_entailment

Tab. 3: Beispiele aus dem QNLI-Datensatz des GLUE-Benchmarks[257]

Im ersten Beispiel enthält der Satz die Antwort auf die Frage, die in diesem Fall „Chelate“ oder „Cetate“ lautet, was den Namen der Stadt bezeichnet, in der die historische Schlacht stattfand. Daher wird das Label „entailment“ vergeben. Im zweiten Beispiel liefert der Satz keine Antwort auf die Frage nach dem Namen des gesuchten Flughafens. Stattdessen thematisiert es die Produktion von Fasern in Neuseeland. Folglich erhält dieses Frage-Antwort-Paar das Label „not_entailment“.

Das Konzept von „entailment“ ist im Bereich der Fragebeantwortung von Bedeutung, da es Parallelen zur Suche nach einer Passage zieht, die eine Frage beantwortet.[258] Darüber hinaus ist der QNLI-Datensatz mit seiner umfangreichen Sammlung kategorisierter Daten öffentlich auf der Platform „Hugging Face“ zugänglich[259] und die zugehörige Originalpublikation wurde laut Google Scholar bereits 6412-mal zitiert.[260]

[257] Auszug aus dem QNLI-Datensatz verfügbar auf: https://huggingface.co/datasets/nyu-mll/glue
[258] Vgl. Paramasivam, Nirmala 2022, S. 9645 ff.
[259] Vgl. Talman et al. 2022, S. 3
[260] Google Scholar: scholar.google.de (Zugriff vom 03.05.2024)

3 Zielspezifikation und Darlegung des Forschungsdesigns

Im vorangegangenen Kapitel wurde der aktuelle Forschungsstand dargelegt sowie eine Übersicht über die für diese Arbeit relevanten Konzepte der Informationsabrufung im Kontext von RAG zur Fragebeantwortung gegeben. Basierend auf diesen Erkenntnissen wird die Hypothese untersucht, dass die Integration von Informationen über die semantische Klassifikation von Fragen und Antworten die Leistungsfähigkeit der Passagenabrufung in einem RAG-System verbessern könnte. Diese Annahme stützt sich auf die historische Verwendung von Klassifikationssystemen (für Fragen) in der Informationsabrufung[261] und die Beobachtung, dass aktuelle Systeme keine Perfekte Abrufungsgenauigkeit vorweisen und direkte lexikalische Ähnlichkeiten oft nicht optimal nutzen können, was eine Kombination mit anderen Relevanzbewertungsmethoden (z.B: lexikalischen Algorithmen) benötigt.[262]

Ausgehend von dieser Hypothese kann nun das Ziel des praktischen Anteils der Arbeit spezifiziert werden. Hierfür wird das Goal-Question-Metric (GQM) Modell von Basili, Caldiera und Rombach (1994) angewandt, das ursprünglich für die Softwareentwicklung konzipiert wurde und eine strukturierte Zerlegung komplexer Problemstellungen in untergeordnete Fragestellungen ermöglicht. Diese Methodik wurde gewählt, weil sie eine präzise und systematische Zieldefinition unterstützt, welcher die Konzeption, Problematik und Perspektive des Forschungsgegenstands integriert. Das GQM-Modell fordert, dass auf konzeptioneller Ebene eine Zielsetzung formuliert wird, welche das Untersuchungsobjekt aus den Blickwinkeln Zweck, Problemstellung und Standpunkt betrachtet. Diese klar definierte Zielsetzung ermöglicht es dann auf operativer Ebene, das Untersuchungsobjekt in Bezug auf das spezifische Problem zu charakterisieren und dessen Leistung aus der gewählten Bewertungsperspektive zu analysieren. Darauf basierend kann auf quantitativer Ebene passende Metriken festgelegt werden, um die zugeordnete Fragestellung zu beantworten.[263]

Im vorliegenden Fall soll im Zuge einer Wissenschaftlichen Arbeit und Unternehmensinterner Entwicklung untersucht werden, ob das derzeitige Abrufungssystem, das den aktuellen Forschungsstand widerspiegelt, durch die Integration von Informationen aus der semantischen Klassifizierung von Fragen und Antworten verbessert werden kann. Dieses Hauptziel wird in zwei spezifische Fragestellungen unterteilt: Erstens, ob eine potenzielle Verbesserung durch ein Re-Ranking-Verfahren erzielt werden kann, und zweitens, ob eine potenzielle Verbesserung durch die vollständige Integration semantischer Klassen in den Abrufungsprozess möglich ist. Die zur Quantifizierung und Beantwortung dieser Fragen verwendeten spezifischen Metriken werden an geeigneter Stelle erörtert. Eine ausführliche Darstellung nach dem GQM-Modell findet sich in Tabelle 4.

---

[261] Vgl. Sarrouti, Ouatik El Alaoui 2017, S. 97 ff.
[262] Vgl. Wang, Zhuang, Zuccon 2021, S. 322 ff.
[263] Vgl. Basili, Caldiera, Rombach 1994, S. 3 f.

Ziel	**Zweck:** Untersuchung der potentiellen Verbesserung (durch Einbindung von semantischer Klassifizierung von Fragen und Passagen: TREC-10, QNLI) **Problem:** der Abrufungsgenauigkeit **Objekt:** in der Abrufungskomponente eines RAG-Systems (IBM Deepsearch) **Standpunkt:** aus der Sicht der internen Entwicklung
Frage 1	Kann die Genauigkeit durch den Einsatz semantischer Klassifizierungsinformationen von Fragen und Passagen in einem Re-Ranking-Verfahren verbessert werden?
Metriken	Siehe Kapitel 4.5.1
Frage 2	Kann die Genauigkeit durch den Einsatz semantischer Klassifizierungsinformationen von Fragen und Passagen in einem integrierten Abrufverfahren verbessert werden?
Metriken	Siehe Kapitel 4.6.1

Tab. 4: Angewendetes GQM-Modell nach Basili, Caldiera, Rombach (1994)

Die präzise definierte Zielsetzung ermöglicht die darauf aufbauende Konzeption des Forschungsdesigns. Für den praktischen Teil dieser Arbeit wird die Design Science Research (DSR) Methode nach Peffers et al. (2007) angewandt, um eine strukturierte Herangehensweise zu gewährleisten. Diese Methode eignet sich besonders für Projekte, welche die Entwicklung eines Softwareprototyps zum Ziel haben,[264] was ideal ist, da für die Untersuchung der potenziellen Verbesserungen des RAG-Systems kein vollständig implementiertes Abrufungssystem für eine Produktivumgebung erforderlich ist. Stattdessen genügt eine prototypische Implementierung, um die notwendigen Erkenntnisse zu gewinnen. Der iterative Prozess der DSR, gekennzeichnet durch die Phasen der Lösungsfindung, des Designs und der Entwicklung, Demonstration sowie der Evaluation,[265] bietet erhebliche Vorteile. Dieser Prozess ermöglicht es, in verschiedenen Iterationen spezifische Lösungen für die jeweiligen Unterfragen (siehe: Tab. 4) zu implementieren und zu evaluieren, sowie verschiedene Konfigurationen des modifizierten Abrufungssystems iterativ zu testen, um die festgelegten Zielsetzungen effektiv zu erreichen. In der vorliegenden Arbeit wird DSR als übergeordneten Forschungsprozess in Form einer Meta-Methodik verwendet,[266] wobei spezifische Methoden zur Erfüllung der Prozessschritte an relevanten Stellenerläutert werden. Die detaillierte Implementierung von DSR im praktischen Teil wird in Kapitel 4.1 ausführlich beschrieben.

[264] Vgl. Holzweißig, S. 29
[265] Vgl. Peffers et al. 2007, S. 54
[266] Vgl. Holzweißig, S. 29

4 Integration semantischer Klassifizierung in ein RAG-Abrufungssystem

In diesem Kapitel liegt der Fokus auf der praktischen Untersuchung der in Kapitel 3 definierten Zielsetzungen und Forschungsfragen. Kapitel 4.1 erläutert die konkrete Anwendung der Forschungsmethodik, die auf dem zuvor definierten Forschungsdesign basiert. Kapitel 4.2 präsentiert die vorbereitenden Iterationen, in denen die Klassifikationssysteme entwickelt werden, die die technische Grundlage für die Integration der semantischen Klassifizierung in die Abrufungsmethodik bilden. Kapitel 4.3 beschreibt das Untersuchungsobjekt dieser Arbeit, das Abrufungssystem. Kapitel 4.4 widmet sich der Vorstellung der Abrufungsdatensätze, anhand derer die in den Hauptiterationen in Kapitel 4.5 und 4.6 implementierten Abrufungsmethoden evaluiert werden. Dabei untersucht Kapitel 4.5 ob ein Re-Ranking-Verfahren mit semantischer Klassifizierung der Fragen und Passagen die Abrufungsgenauigkeit verbessern kann. Kapitel 4.6 untersucht, ob die genannte Verbesserung durch eine enge Integration der semantischen Klassifizierung in die Abrufungsmethodik erreicht werden kann.

Der vollständige Quelltext dieser Arbeit ist im dafür angelegten GitHub-Repository zu finden. Die Ordnerstruktur ist an die Kapitelstruktur dieser Arbeit angelehnt, und die Dateien sind entsprechend ihrer Funktion benannt.[267]

4.1 Zielsetzung und Forschungemethodik

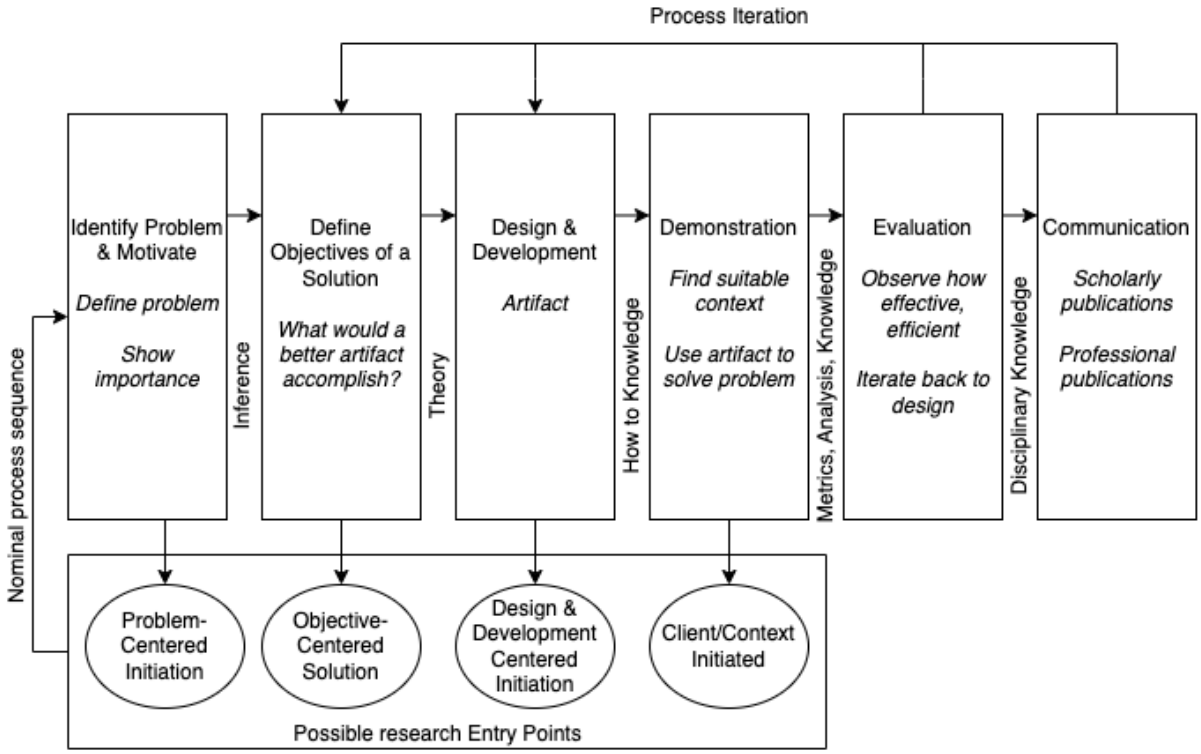


Abb. 4: DSR-Forschungsprozess nach Pfeffers et al. (2007)[268]

[267] Siehe: https://github.com/TienDeeLInPrivate/retireval_with_semantic_classes
[268] Vgl. Peffers et al. 2007, S. 54

Basierend auf der zuvor definierten Zielsetzung sowie der Darlegung des Forschungsdesigns wird in diesem Kapitel die Anwendung der DSR-Methodik formuliert und dokumentiert. Dies dient dazu, das Vorgehen transparent und für Dritte reproduzierbar zu machen. Der DSR-Forschungsprozess nach Peffers et al. (2007), welcher bei Bedarf wiederholte Rückiterationen erlaubt,[269] wird in Abbildung 4 vorgestellt.

In dieser Arbeit beginnt der Einstieg in der Phase „Define Objectives as a Solution", passend zur zielorientierten Lösungsausrichtung des Forschungsbeitrags. Entsprechend der Verwendung von DSR als Metamethodik wurde das Ziel bereits in Kapitel 3 gemäß dem GQM-Modell definiert. Die weiteren Schritte des DSR-Iterationsprozesses sind wie folgt strukturiert: In der Phase „Design & Development" wird das RAG-System unter Einbeziehung der semantischen Klassifizierung implementiert. In der Demonstrationsphase wird das entwickelte System der aktuellen Iteration auf den spezifischen Anwendungsfall und die zu abrufenden Daten angewendet. In der Evaluationsphase wird die Leistung des modifizierten Systems bewertet und es wird diskutiert, ob weitere Iterationen erforderlich sind. Zudem dienen die verschiedenen Iterationen dazu, unterschiedliche Konfigurationen des Systems zu testen. Die Evaluation wird unter Laborbedingungen durchgeführt, um eine kontrollierte Umgebung zu gewährleisten, in der Kausalzusammenhänge genau untersucht werden können. Diese künstliche Umgebung ermöglicht die Manipulation experimenteller Variablen in wiederholbaren Szenarien, um präzise und reproduzierbare Ergebnisse zu erzielen.[270] Diese sind essenziell, um die definierten Zielsetzungen und Forschungsfragen zuverlässig zu beantworten.

4.2 Vorbereitende Iterationen: Entwicklung der Klassifikationssysteme

Zur Erreichung des definierten Ziels ist die Entwicklung von Klassifikationsmodellen zur semantischen Klassifikation von Fragen und Passagen erforderlich. Die Forschenden Li und Roth konnten ihre eigene TREC-10 Taxonomie mithilfe eines Klassifikationsmodells basierend auf der SNoW-Architektur mit einer Genauigkeit von 98,8% in den Hauptkategorien klassifizieren.[271] Das derzeit führende Team in der öffentlichen Rangliste des GLUE-Benchmarks erzielt mit dem Turing ULR-v6 Sprachmodell von Microsoft eine Genauigkeit von 96,7% auf dem QNLI-Datensatz.[272] Diese Leistungen sind bemerkenswert, jedoch ist es im zeitlichen und ressourcenbedingten Rahmen dieser Arbeit nicht möglich, die Modelle und Systeme dieser Forschenden zu replizieren. Zudem stehen keine detaillierten Implementierungsdetails oder direkt zugängliche Quellcodes dieser Systeme zur Verfügung, die eine Nachbildung in der vorgege-

[269] Vgl. Peffers et al. 2007, S. 56
[270] Vgl. Wilde, Hess 2007, S. 282
[271] Vgl. Li, Roth 2002, S. 4 ff.
[272] GLUE 2024. GLUE Benchmark: https://gluebenchmark.com/leaderboard (Zugriff vom 15.04.2024)

benen Zeit erlauben würden. Daher besteht das Ziel der ersten DSR-Iterationen dieses Forschungsprojekts darin, Klassifikationsmodelle für die TREC-10 und GLUE-QNLI Klassen zu entwickeln, die erstens eine angemessene Genauigkeit erreichen und zweitens innerhalb des kurzen Zeitrahmens umsetzbar sind.

Angesichts der zuvor dargestellten Umstände wird in dieser Arbeit der Einsatz vortrainierter generativer Sprachmodelle in Verbindung mit Prompt-Engineering verfolgt, um die erforderlichen Klassifikationen durchzuführen. Dieser Ansatz ist wissenschaftlich etabliert, da bereits demonstriert wurde, dass vortrainierte Modelle fähig sind, aus einer kleinen Anzahl von Beispielen im Prompt zu lernen und in der Verarbeitung natürlicher Sprache zuverlässig zu agieren.[273] Ferner nutzt der gegenwärtige Führende der GLUE-Benchmark-Rangliste, das Turing ULR-v6 Modell, welches ebenfalls ein vortrainiertes Sprachmodell ist.[274] Diese Methode ist besonders vorteilhaft, da sie die schnelle Entwicklung eines funktionsfähigen Klassifikationsmodells ermöglicht. Zudem besteht die interne Anweisung des Unternehmens, ausschließlich die eigene Plattform IBM Research Big AI Model (BAM) Laboratory zu verwenden, die generative Sprachmodelle sowohl über API als auch eine Web-UI bereitstellt. Daher ist die Auswahl der Sprachmodelle auf diejenigen beschränkt, die auf BAM verfügbar sind.

4.2.1 1. Iteration: Klassifikationssystem für GLUE-QNLI

Im Rahmen des DSR-Prozesses zielt diese Iteration darauf ab, ein System zu entwickeln, das mit einem vortrainierten Sprachmodell den QNLI-Datensatz des GLUE-Benchmarks klassifizieren kann. Dies soll mit einem vertretbaren zeitlichen Aufwand und einer angemessenen Genauigkeit erreicht werden. Für die Implementierung wird der Datensatz **nyu-mll/glue** verwendet, der öffentlich auf der Plattform „Hugging Face“ zugänglich ist.[275] Der QNLI-Teil dieses Datensatzes setzt sich aus Fragen, Sätzen und zugehörigen Labels zusammen und umfasst drei Segmente: 104.743 Zeilen mit annotierten Trainingsdaten, 5.463 Zeilen mit annotierten Validierungsdaten und 5.463 Zeilen mit nicht annotierten Testdaten. Aufgrund der fehlenden Annotationen bei den Testdaten wird stattdessen der Validierungsdatensatz als Testset verwendet. Wie bereits in Kapitel 2.4 erläutert, erhalten die Zeilen das Label „entailment“, wenn

[273] Vgl. Sun et al. 2023, S. 1
[274] Vgl. Microsoft Bing Blogs 2022. Microsoft Turing Universal Language Representation model, T-ULRv6, tops both XTREME and GLUE leaderboards with a single model: https://blogs.bing.com/search-quality-insights/october-2022/Microsoft-Turing-Universal-Language-Representation-model,-T-ULRv6,-tops-both-XTREME-and-GLUE-leaderb (Zugriff vom 15.04.2024)
[275] Daten öffentlich verfügbar auf: https://huggingface.co/datasets/nyu-mll/glue (Zugriff vom 17.04.2022)

der Satz die Antwort auf die Frage enthält, und „not_entailment“, wenn dies nicht der Fall ist. Abbildung 5 stellt die entsprechende Verteilung der Labels im Testdatensatz dar.

Die Visualisierung verdeutlicht eine ausgewogene Verteilung der Testdaten, was bedeutet, dass das zu entwickelnde Modell daraufhin geprüft wird, ob es beide Labels zuverlässig vorhersagen kann. Da der Datensatz Inhalte aus verschiedenen Domänen und inhaltlichen Bereichen umfasst, lässt sich vermuten, dass das Klassifikationssystem daraufhin getestet wird, das Konzept des „entailment“ domänenübergreifend korrekt anzuwenden.

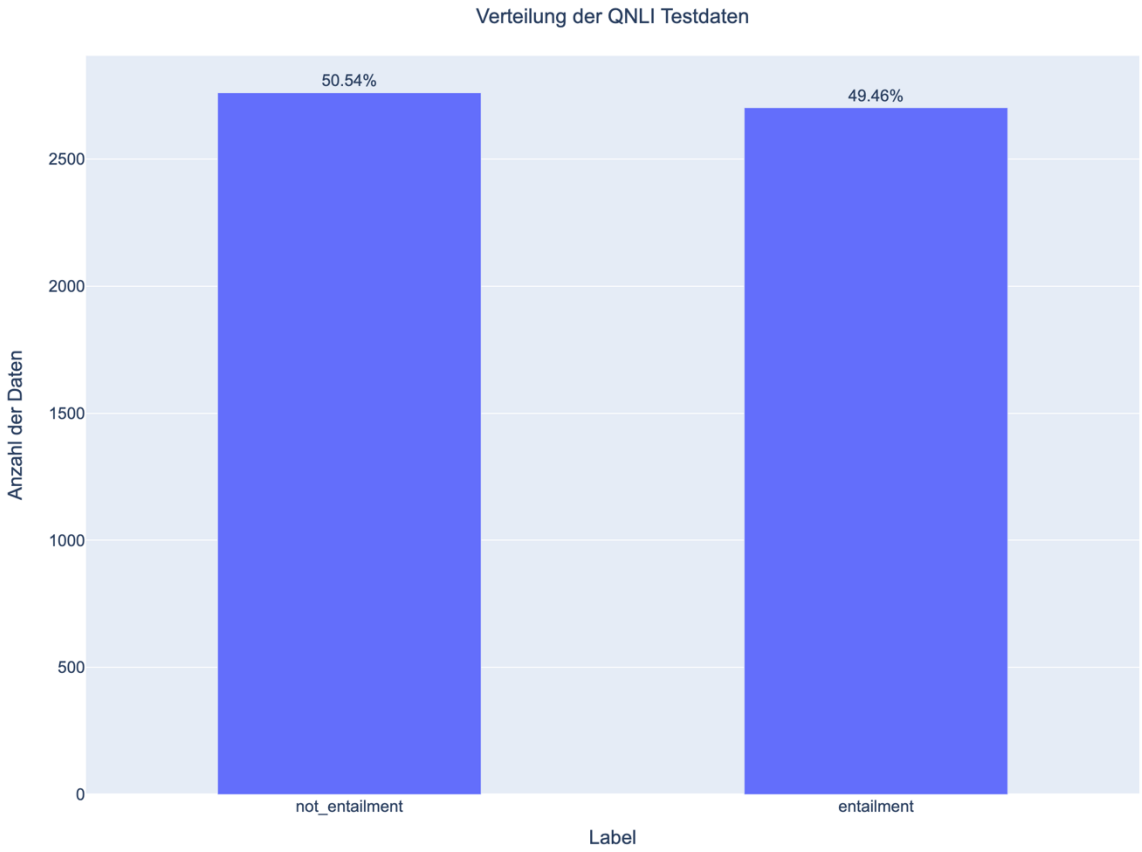


Abb. 5: Visualisierung der QNLI-Testdaten

Entwicklungsphase

Nach explorativen Tests in der webbasierten Benutzeroberfläche von BAM wurde das Modell „google/flan-ul2“ zur Durchführung der Klassifikationsaufgabe ausgewählt, da es auf Anweisungen dieser Art am effektivsten reagierte. Dieses Encoder-Decoder-Modell, entwickelt von Google, basiert auf der T5-Architektur.[276] Für die Formulierung der Anweisungen an das Modell wird die Few-Shot-Prompting-Methodik nach Brown et al. (2020) eingesetzt, bei der zufällig ausgewählte Datenpunkte aus den Trainingsdaten als Beispiele in den Prompt

[276] Für weitere Informationen sei auf die offizielle Dokumentation verwiesen, verfügbar unter: https://huggingface.co/google/flan-ul2 (Zugriff vom 17.04.2023)

eingebunden werden.[277] Die Generierung der Antworten wird gemäß der Methode von Reynolds und McDonell (2021) gesteuert, wobei die zu klassifizierenden Daten analog zu den vorherigen Beispielen im Prompt zuletzt dargestellt und die Position für das Label freigelassen wird.[278] Die Struktur des Prompts gestaltet sich folgendermaßen:

Prompt-Beginn	Classify this question and sentence pair based on its entailment in one of these categories: entailment, not_entailment. The label 'entailment' when the sentence contains the answer to the question. The label is labeled 'not_entailment' when the sentence does not contain the answer to the question. Here are some example questions together with the class label:
Beispiele	Question: Who did NASA recruit by using flawed safety numbers? Sentence: He concluded that the space shuttle reliability estimate by NASA management was fantastically unrealistic, and he was particularly angered that NASA used these figures to recruit Christa McAuliffe into the Teacher-in-Space program. Label: entailment *(Es werden für jedes Label jeweils 10 Beispiele angeführt)*
Zu klassifizierender Datenpunkt	Question: {row['question']} Sentence: {row['sentence']} Label:

Tab. 5: Prompt-Struktur zur Klassifizierung (QNLI)[279]

Es werden zehn Beispiele pro Klasse ausgewählt, da die BAM-Plattform einerseits empfiehlt, insgesamt nicht mehr als 30 Beispiele zu übermitteln, und andererseits, um ausreichend Kontextlänge für das Modell freizuhalten. Dies ist besonders wichtig für den Schritt, in dem das Klassifikationsmodell auf längere Fragen und Passagen angewendet wird. Zudem wird die Temperatur initial auf 0 gesetzt, da diese Aufgabe keine Kreativität erfordert. Beispiele aus dem GLUE-Datensatz werden genutzt, um sicherzustellen, dass die Methode potenziell auf verschiedene Abrufungsdaten in einer Produktivumgebung übertragbar ist. Dieser Ansatz ist besonders relevant, wenn nicht stets neue Datensätze manuell annotiert werden können. Aufgrund seiner thematischen Vielfalt dient der GLUE-Datensatz als repräsentative Grundlage und unterstützt die Generalisierbarkeit des Klassifikationsmodells.

[277] Vgl. Brown et al. 2020, S. 4
[278] Vgl. Reynolds, McDonell 2021, S. 2
[279] Der vollständige Prompt ist in Anhang 1/1 dokumentiert

Demonstration & Evaluation

Der beschriebene Prompt und die Modelleinstellungen wurden erfolgreich auf den definierten Testdatensatz angewandt. Das Modell generierte präzise Labels zu jedem Datenpunkt, ohne Halluzinationen, und erreichte eine Genauigkeit von etwa 94,2%. Dies entspricht einer geringen Abweichung von 2,5% im Vergleich zu den führenden Ergebnissen des GLUE-Benchmarks. Angesichts der zeitlichen Rahmenbedingungen dieser Arbeit wird diese Genauigkeit akzeptiert, und das Modell wird für den Einsatz festgelegt. Der Fokus der Arbeit liegt nicht darauf, die Klassifikationsgenauigkeit zu maximieren, sondern vielmehr darauf, eine robuste Lösung bereitzustellen, ohne eine letzte Optimierung der Genauigkeitsprozente anzustreben, da dies nicht zentral für die Zielerreichung ist. Daher ist keine weitere Iteration erforderlich.

4.2.2 2. Iteration: Klassifikationssystem für TREC-10 Fragen

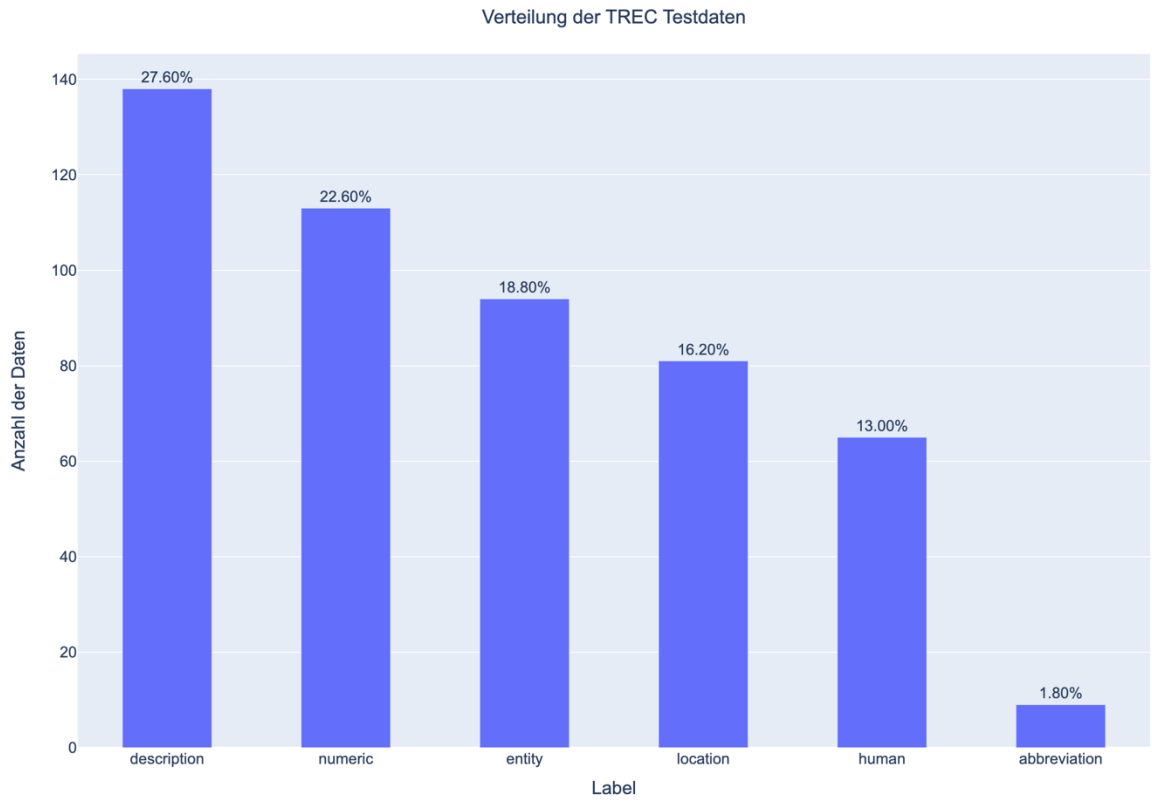


Abb. 6: Visualisierung der TREC-Testdaten

Ähnlich wie in der vorherigen Iteration zielt diese darauf ab, ein Klassifikationssystem für die TREC-10 Fragen zu entwickeln, basierend auf der Taxonomie, die in Kapitel 2.4 vorgestellt wurde. Nach eingehender Betrachtung der Taxonomie wurde entschieden, in dieser Arbeit nur die sechs Hauptkategorien (abbreviation, numeric, human location, description, entity) zu berücksichtigen, da die spezielleren und möglicherweise nicht generalisierbaren Unterkategorien wie “religion”, “desease medicine” oder “sport” eventuell nicht geeignet für den Ziel-Datensatz sind oder auf andere potenzielle Abrufungs-Datensätze übertragbar sind. Durch die Auswahl

der Hauptkategorien wird die Klassifizierungslogik so angepasst, dass auch andere semantische Fragearten unter die Hauptkategorie fallen können, wenn sie logisch dazu passen, aber nicht explizit in einer der spezifischen Unterkategorien aufgeführt sind. Zur Implementierung wird der „trec"-Datensatz verwendet, der auf der Plattform „Hugging Face" verfügbar ist.[280] Dieser umfasst 5450 Zeilen Trainingsdaten und 500 Zeilen Testdaten, die nach der Art der benötigten Antwort annotiert sind. Die Verteilung des Testdatensatzes ist in Abbildung 6 visualisiert.

Die Testdaten zeigen eine ungleichmäßige Verteilung über die verschiedenen Hauptkategorien hinweg. Dies legt die Vermutung nahe, dass das Modell nicht ausgewogen auf seine Fähigkeit hin getestet wird, Fragen über alle Kategorien hinweg gleichmäßig gut zu klassifizieren. Aufgrund des Mangels an alternativen Datensätzen wird dieser Datensatz weiterhin verwendet, allerdings wird dieser Umstand bei der Ergebnisevaluation in Kapitel 5 berücksichtigt.

Entwicklungsphase

Aufgrund der zufriedenstellenden Ergebnisse der vorherigen Iteration wird in dieser Iteration der gleiche Ansatz verfolgt. Dabei wird erneut das Modell „google/flan-ul2" unter Anwendung derselben Few-Shot-Prompting-Methode eingesetzt, um die Fragen zu klassifizieren. Die Struktur des Prompts gestaltet sich wie folgt:

Prompt-Beginn	Classify this question based on its intent in one of these categories: abbreviation, entity, description, human, location, or numeric. Focus on the interrogative pronouns. The result of a query can only consist of a single word. In Example: description Here are some example questions together with the class label of the question:
Beispiele	Question: What state is known as the Hawkeye State ? Label: location *(Es werden für jedes Label jeweils 5 Beispiele angeführt)*
Zu klassifizierender Datenpunkt	Question: {row['text']} Label:

Tab. 6: Prompt-Struktur zur Klassifizierung (TREC)[281]

Wie in der vorherigen Iteration wurden die Beispiele zufällig aus den Trainingsdaten ausgewählt. Die Begründung hierfür bleibt dieselbe wie zuvor. Aufgrund der genannten Einschränkungen der BAM-Plattform wurde die Anzahl der Beispiele auf insgesamt 30 festgelegt. In

[280] Daten öffentlich verfügbar auf: https://huggingface.co/datasets/trec (Zugriff vom 17.04.2024)
[281] Der vollständige Prompt ist in Anhang 1/2 dokumentiert

diesem Fall können mehr Beispiele verwendet werden, da die Klassifizierung – anders als beim QNLI-Datensatz – nur die Frage selbst betrifft und Sätze vorhanden sind. Auch bei dieser Aufgabe ist keine Kreativität erforderlich, weshalb die Temperatur auf 0 gesetzt wird.

Demonstration & Evaluation

Auch in diesem Fall wurde der Prompt zusammen mit den Modelleinstellungen erfolgreich auf den Testdatensatz angewendet. Das Modell erreichte eine Genauigkeit von 97%, ohne Anzeichen von Halluzinationen. Die Differenz dieser Klassifizierungsgenauigkeit der Hauptkategorien zu der Klassifizierungsmethode von Li & Roth (2002) beträgt 1,8%.[282] Angesichts der zeitlichen Rahmenbedingungen und des Schwerpunkts dieser Arbeit wird diese Genauigkeit akzeptiert und das Modell für den Einsatz festgelegt. Wie auch in der letzten Iteration, ist eine Optimierung der letzten Prozentpunkte der Genauigkeit nicht entscheidend für die Zielerreichung, daher wird keine weitere Iteration benötigt.

4.2.3 3. Iteration Klassifikationssystem für Passagen

Die Nutzung der QNLI-Klassen für die Abrufung ist intuitiv, da einfach bewertet werden kann, welche Passagen die Antwort auf die gestellte Frage enthalten („entailment“) und diese entsprechend höher in der Relevanz eingestuft werden. Die Anwendung der sechs groben TREC-Klassen ist jedoch nicht direkt möglich, da diese Taxonomie sich ausschließlich auf die Fragen beziehen. In dieser Arbeit wird ein Ansatz verfolgt, der prüft, ob Passagen die gesuchten Antworttypen oder Informationstypen enthalten. Dabei werden die sechs Frageklassen auf die Passagen angewandt, um zu beschreiben, welche Art von Informationen die Passage liefert. Beispielsweise könnte eine Passage numerische Informationen liefern, indem sie ein Datum enthält (Label: „numeric“), oder deskriptive Informationen beinhalten, indem sie eine Beschreibung von etwas enthält (Label: „description“). Für diesen Zweck wird ein Klassifikationsmodell benötigt, das Textpassagen in die sechs groben TREC-Klassen einteilen kann. Aufgrund der positiven Ergebnisse der vorherigen zwei Iterationen wird auch hier die Methode des Few-Shot Prompting mit dem Modell google/flan-ul2 angewendet.

Bei der Entwicklung des geforderten Klassifikationssystems stehen zwei wesentliche Herausforderungen im Fokus: Erstens stehen derzeit keine Daten zur Verfügung, die als Beispiele für das Few-Shot-Prompting oder als Evaluationsdatensatz zur Überprüfung der Klassifizierungsmethode verwendet werden könnten. Zweitens ist eine eindeutige Klassifizierung einer Passage in eine einzelne Kategorie nicht immer möglich. Zum Beispiel enthält die Passage: „Die Duale Hochschule Baden-Württemberg (DHBW) ist eine staatliche duale Hochschule Deutschlands, die an ihren neun Standorten 34 akkreditierte Bachelor-Studiengänge in verschiedenen Bereichen anbietet. Dabei wird der Stuttgarter Standort in der Paulinenstraße 50 von einigen Studierenden als der schönste bezeichnet.“, mehrere Informationstypen: „abbreviation“

[282] Vgl. Li, Roth 2002, S. 5

(DHBW), „numeric“ (neun Standorte, 34 Studiengänge), „location“ (Paulinenstraße 50) und „description“ (Beschreibung der DHBW).

Um der ersten Herausforderung zu begegnen, werden manuell annotierte Passagen aus dem intern von der Abteilung erstellten Wikipedia-Datensatz als Beispiele für das Few-Shot-Prompting verwendet. Dieser Datensatz, der in Kapitel 4.4 detaillierter beschrieben wird, umfasst Inhalte aus unterschiedlichsten Domänen und Darstellungsstrukturen. Es wird angenommen, dass die Verwendung dieser Beispiele eine gewisse Generalisierbarkeit des Modells ermöglicht.

Es hat sich bei explorativem Prompt-Tuning in der Deepsearch-UI gezeigt, dass das google/flan-ul2 Modell Schwierigkeiten hat, den Prompt zur Multi-Klassen-Klassifikation, insbesondere wenn mehrere Klassen pro Datenpunkt möglich sind, zuverlässig zu verstehen und umzusetzen. Daher wird das "One-vs-All" (OVA) Dekompositionsschema nach Rifkin und Klautau (2004) angewandt, um die Aufgabe in einzelne binäre Klassifikationsaufgaben zu zerlegen.[283] Bereits in Kapitel 4.2.1 wurde festgestellt, dass das google/flan-ul2 Modell bei der binären Klassifikation zuverlässige Ergebnisse liefert. Nach der OVA-Dekomposition wurde die Klassifikationsaufgabe in 6 separate binäre Klassifikationen unterteilt, wobei jede binäre Klassifikation entscheidet, ob eine Passage Informationen einer bestimmten Kategorie, wie z.B. „abbreviation“ für Abkürzungen, enthält oder nicht („not_abbreviation“).[284] Die OVA-Methodik spiegelt sich darin wider, dass eine Passage nur dann positiv klassifiziert wird, wenn sie genau zu der betreffenden Kategorie passt und alle anderen möglichen Kategorien als negativ („not_...“) eingestuft werden.[285] Nachfolgend wird eine beispielhafte Struktur des Prompts präsentiert, die Tabelle wird auf der nächsten Seite weitergeführt.

Prompt-Beginn	Classify the targeted text-passage based on whether it provides information in the form of descriptions or not. If it contains information in the form of descriptions, classify it as 'description' If it does not contain information in the form of descriptions, classify it as 'not_description' Answer by providing only the Label and nothing else. Here are some examples with a Text-Passage followed by its label:
Beispiele	Text-Passage: *(1 Beispielpassage aus Wikipedia-Daten)* Label: description *(….)* Text-Passage: *(1 Beispielpassage aus Wikipedia-Daten)*

[283] Vgl. Rifkin, Klautau, S. 1
[284] Vgl. Galar et al. 2011, S. 1763
[285] Vgl. Rifkin, Klautau, S. 1 f.

	Label: not_description *(Es werden für jedes Label jeweils 2 Beispiele angeführt)*
Zu klassifizierender Datenpunkt	Target Text-Passage: {row['context']} Label:

Tab. 7: Beispiel der Prompt-Struktur für eine binäre Klassifizierung

Demonstration & Evaluation:

Das Klassifikationssystem mit sechs spezifischen Prompts (eines für jede binäre Klassifikation) wurde mit einer Temperatur von 0 und unter Einsatz des Modells google/lan-ul2 auf die Passagen des Wikipedia-Datensatzes angewendet. Das Ergebnis zeigte keine Halluzinationen seitens des Sprachmodells, und es gab lediglich drei Fälle, in denen die zu klassifizierende Passage zusammen mit dem Prompt die maximale Kontextlänge des Modells überschritt (“error”). Die Verteilung, dargestellt in der nachfolgenden Abbildung 7, erscheint auf den ersten Blick plausibel und zeigt die Einteilung der Passagen in die verschiedenen Informationskategorien. Mit dem Mangel zuverlässiger Testdaten wird die Annahme getroffen, dass das Klassifikationssystem die Passagen mit angemessener Genauigkeit einordnet, und das Ergebnis dieser Iteration wird für die Verwendung in weiteren Iterationen akzeptiert.

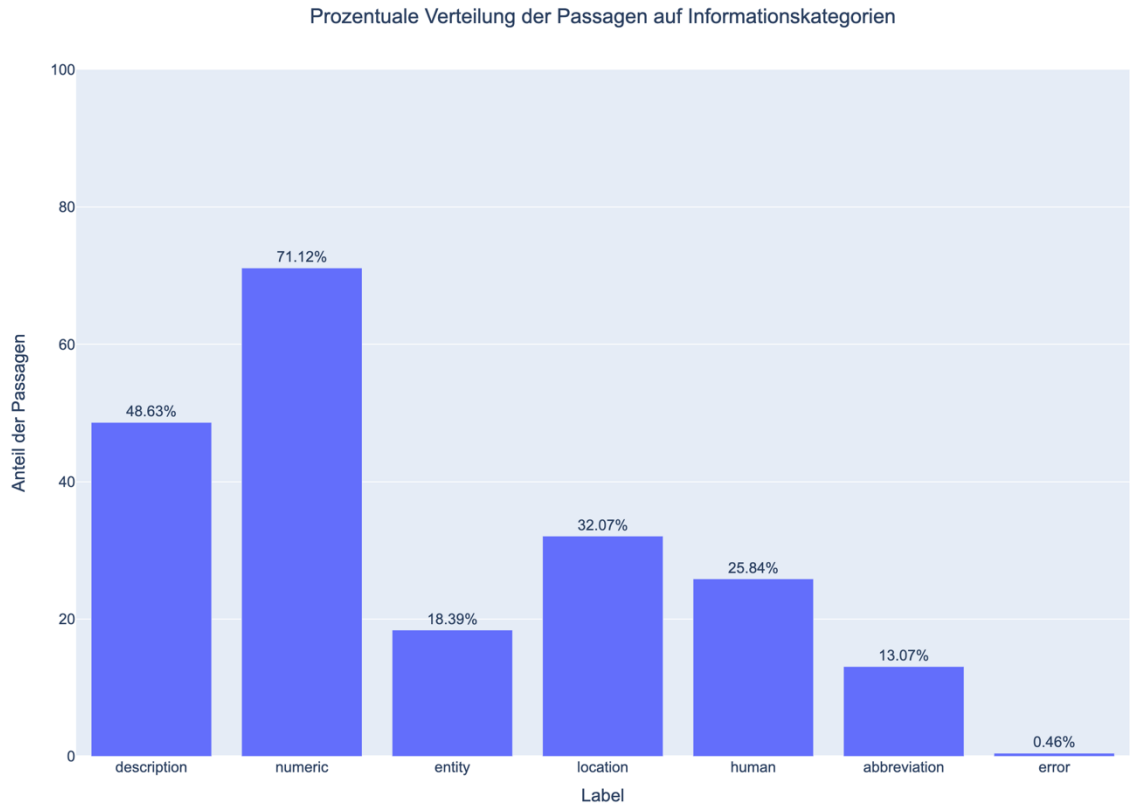


Abb. 7: Verteilung der Wikipedia-Passagen auf TREC-Klassen

4.3 Vorstellung des Abrufungssystems

Das RAG-System von IBM, bekannt unter dem Produktnamen IBM Deepsearch, bildet den Forschungsgegenstand dieser Arbeit, auf den alle Experimente angewendet werden. Das Deepsearch-RAG ist ein naives RAG-System, das die Komponenten Indexierung, Eingabe, Abruf und Generierung umfasst,[286] wie in Kapitel 2 erläutert (siehe Abb. 1). Im Einklang mit dem Schwerpunkt dieser Arbeit wird das Abrufsystem von IBM Deepsearch im Folgenden detailliert beschrieben:

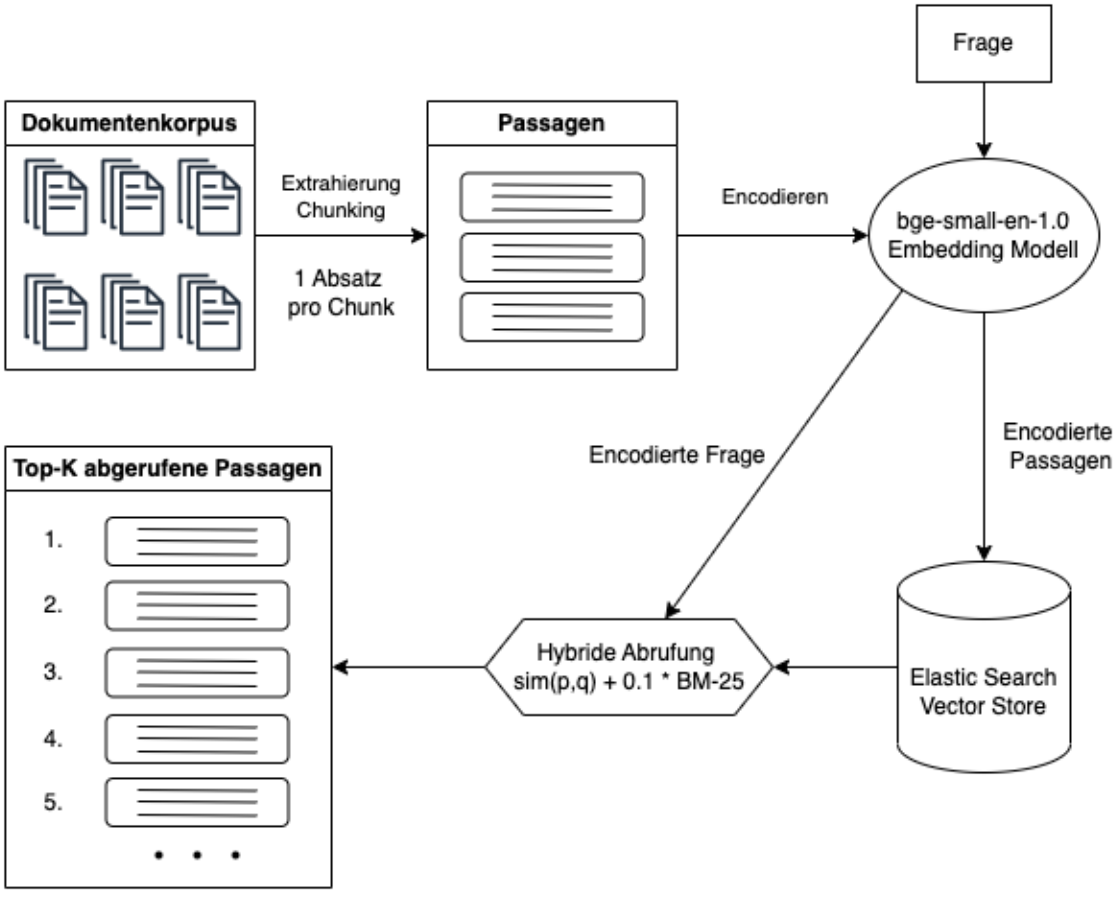


Abb. 8: Abrufungssystem von IBM Deepsearch

Im Abrufungsprozess wird zunächst der Dokumentenkorpus eingelesen und die Inhalte in Chunks unterteilt, wobei ein Absatz eines Dokuments einen Chunk bildet. Diese Chunks variieren somit in ihrer Größe. Jeder Chunk enthält die Text-Passage und Informationen darüber,

[286] Vgl. Gao et al. 2024, S. 3 f.

zu welchem Dokument und welcher spezifischen Seite es gehört und an welcher Position der Absatz steht. Die reinen Textpassagen werden anschließend mit dem Modell „bge-small-en“ (Version 1.0) von BAAI[287] in dichte Vektorrepräsentationen umgewandelt. Diese encodierten Passagen werden in einem „Elasticsearch“-Vektorspeichersystem gespeichert. Bei der Eingabe einer Frage durch den Anwender wird diese ebenfalls mit dem „bge-small-en“ Modell encodiert und der Abrufungsprozess gestartet. Der Abrufungsalgorithmus arbeitet nach einem hybriden Ansatz und kombiniert die semantische Ähnlichkeit mit dem BM-25-Wert von Fragen und Passagen durch eine lineare Kombination.[288] Dabei wird die semantische Ähnlichkeit zum mit 0,1 gewichteten BM-25-Wert addiert. Basierend auf diesem aggregierten Relevanzwert werden die Passagen entsprechend ihrer Bedeutung für die Frage sortiert. Das System ermöglicht es, intern festzulegen, wie viele Passagen abgerufen werden sollen.

Das Abrufungssystem von IBM Deepsearch repräsentiert den aktuellen Forschungsstand und nutzt die in der Wissenschaft diskutierte und angewandte hybride Abrufmethode. Der Interpolationsparameter (0,1) wurde basierend auf internen Optimierungsstudien festgelegt.

4.4 Vorstellung der Abrufungsdaten

In dieser Arbeit werden zwei Datensätze verwendet, um das Abrufungssystem zu evaluieren. Im Folgenden werden diese Abrufungsdaten vorgestellt:

Wikipedia-Datensatz

Intern wurde dieser Datensatz erstellt, um die Leistung des Abrufungssystems auf domänenoffenen Daten zu testen. Eine Vielzahl von zufällig gewählten Wikipedia-Artikeln wurde über die Wikipedia-API[289] bezogen und gemäß Abbildung 7 in Chunks und Passagen unterteilt, encodiert und im Vektorspeichersystem gespeichert. Eine Kopie der Originalpassagen wurde zur synthetischen Fragegenerierung genutzt. Hierfür wurden die Passagen nacheinander einem generativen Sprachmodell der Mistral-Familie zugeführt, mit der Anweisung, passende Fragen zu erzeugen, die mit der jeweiligen Passage beantwortet werden können. Für dieselben Passagen wurden mehrere unterschiedliche Fragen generiert. Dadurch entstand ein Datensatz aus 3256 Frage-Passagen-Paaren, die zur Evaluierung des Abrufungssystems herangezogen werden können. Da jede Frage eine Ursprungspassage hat, wird die Passage, aus der die Frage generiert wurde, als die einzig korrekte Passage angesehen, mit der die Frage beantwortet werden kann. Eine besondere Herausforderung dieses Datensatzes ist die Tat-

[287] Öffentlich verfügbar und dokumentiert auf: https://huggingface.co/BAAI/bge-small-en (Zugriff vom 19.04.2024)
[288] Ausführlich erörtert in Kapitel 2.3.2
[289] Eine Dokumentation dieser API findet sich unter: https://www.mediawiki.org/wiki/API:Main_page/de (Zugriff vom 20.04.2024)

sache, dass es neben reinen Textpassagen auch Auflistungen und in Text umgewandelte Tabellen enthält. Die Erfassung der semantischen Bedeutung von Text in diesen unterschiedlichen Darstellungsformen könnte eine Herausforderung für das Abrufungssystem darstellen.

Diese Herausforderung reflektiert die Praxis, da Informationen oft in verschiedenen Strukturen wie Tabellen und Listen auftreten, die in der Anwendung häufig vorkommen. Es ist realistisch, dass Nutzer des Systems Fragen zu spezifischen Dokumenten stellen, die möglicherweise nur durch eine Tabelle beantwortet werden können. Daher ist es wichtig, die Leistung des Abrufungssystems auch für diese Informationsstrukturen zu bewerten.

```
|--------------|-----------------------------------------------------------|
| Born         | George Joseph Laurer III September 23, 1925 Manhattan, New |
| Died         | December 5, 2019 (aged 94) Wendell, North Carolina, U.S.   |
| Alma mater   | University of Maryland                                    |
| Notable work | Universal Product Code                                    |
```

Abb. 9: Beispiel einer in Text umgewandelten Tabelle

Abbildung 8 präsentiert eine Tabelle, die aus einem Wikipedia-Artikel des Datensatzes extrahiert und in Textform umgewandelt wurde. Diese Tabelle wird als eine Passage angesehen. Eine der generierten Fragen zu dieser Passage lautet: „What is George Joseph Laurer III most famous for?“

```
DS8888F
Dual 48-core POWER8-based controllers
Running SMT-4 for 192 threads
Up to 2 TiB Cache
High Performance Flash Enclosure: integrates and optimizes flash technology in the DS8888F
Can contain up to 480 1.8 ' flash cards in the High-Performance Flash Enclosure (HPFE)
Up to 16 Flash Enclosures per System : 192 TB raw per system
DS89#0F-released in 2020 [14]
IBM DS8910F $^{[15]}$-Rack-mounting (20U, 19U without KVM)
based on a dual IBM Power Systems S922, S914, or S924 controllers
IBM DS8950F
42U assembled rack cabinet
```

Abb. 10: Beispiel einer reinen Auflistung

Abbildung 9 präsentiert eine reine Auflistung, deren Informationsstruktur auf den ersten Blick nicht sofort erkennbar ist. Eine der generierten Fragen zu dieser Abbildung lautet: “How many flash cards can the High-Performance Flash Enclosure (HPFE) of the DS8888F hold?”

PubMed-Datensatz[290]

Der zweite Datensatz basiert auf Auszügen aus PubMed Central, einem Teil der von der "National Library of Medicine" bereitgestellten PubMed-Datenbank, die speziell für die Suche und Abfrage medizinbiologischer Fachliteratur dient. PubMed Central umfasst vollständige Artikel der PubMed-Datenbank, die öffentlich zugänglich sind. Für die Evaluierung des Abrufungssystems wurden speziell Artikel zum Thema Herzklappenchirugie (engl. valve replacement) ausgewählt, die durch eine Schlagwortsuche identifiziert wurden. Wie beim Wikipedia-Datensatz wurden auch diese Artikel in Chunks unterteilt, encodiert und im Vektorspeichersystem gespeichert. Die Methodik zur synthetischen Fragegenerierung wurde ebenfalls angewandt. Dieser Datensatz, der ebenfalls Texte in verschiedenen Informationsstrukturen wie Tabellen und Auflistungen enthält, ist speziell darauf ausgelegt, die Leistung des Abrufungssystems auf domänenspezifische Daten zu prüfen. Es umfasst genau 10.000 Fragen und Passagenpaare.

4.5 Untersuchung der potentiellen Verbesserung durch ein Re-Ranking verfahren

Mit der abgeschlossenen Entwicklung der Klassifikationssysteme für die semantischen Klassen werden in den nachfolgenden Iterationen der Fokus darauf gelegt Frage 1 (siehe: Tab. 4) zu beantworten, nämlich ob es möglich ist durch die Einbindung der vorgestellten semantischen Klassen in einem Re-Ranking Verfahren das Abrufungssystem zu verbessern.

4.5.1 1. Iteration: QNLI-Klassen im Re-Ranking Verfahren

Zu Beginn wird das Klassifikationssystem, das in Kapitel 4.2.1 entwickelt wurde, eingesetzt, um die Passagen mithilfe der QNLI-Klassen zu ordnen. Nach der Abrufung werden die Passagen erneut sortiert, indem festgestellt wird, ob sie die Antwort auf die gestellte Frage enthalten, das heißt, ob sie zusammen mit der Frage das Label „entailment“ erhalten.

Entwicklungsphase

Der Ausgangspunkt dieser Iteration ist die Ausgabe des untersuchten Abrufungssystems, wie in Kapitel 4.3 beschrieben. Aufgrund der begrenzten verfügbaren Rechenleistung werden lediglich die Top-10 von IBM Deepsearch abgerufenen Passagen berücksichtigt, die hinsichtlich des Labels „entailment“ neu nach Relevanz sortiert werden sollen. Die Entscheidung, nur die Top-10 Passagen zu betrachten, basiert darauf, dass die Klassifikation von 3000 bis 10000 Fragen mit jeweils 10 abgerufenen Passagen pro Frage, wenn jede Klassifikation eine Sekunde benötigt, bereits viele Stunden in Anspruch nehmen würde. Hinzu kommt, dass die verwendete BAM-API eine maximale Nutzungsdauer von 24 Stunden aufweist. Die Fokussierung auf die Top-10 Passagen ist ebenfalls ein in der wissenschaftlichen Literatur angewandtes Vorgehen.[291] Zudem würde eine zu große Menge an Passagen (im Kontext vom naiven RAG)

[290] Aufgrund unklarer Lizenzbestimmungen bei einigen Dokumenten dieser Kollektion können diese Daten in der vorliegenden Arbeit nicht gezeigt werden (Abteilungsinterne Anweisung)
[291] Vgl. Prakash, Killingback, Zamani 2021, S. 1730

aufgrund der begrenzten Kontextlänge des antwortgenerierenden Sprachmodells nicht in die Kontextlänge passen und könnte somit nicht zur Antwortgenerierung herangezogen werden.

Die Neuordnung der abgerufenen Passagen erfolgt, indem alle Passagen, die zusammen mit der Frage als „entailment“ klassifiziert wurden, in der Rangordnung vor jenen Passagen platziert werden, die als „not_entailment“ klassifiziert wurden. Dabei werden „entailment“-Passagen nicht vor anderen „entailment“-Passagen platziert, die in der ursprünglichen oder aktuellen Rangordnung vor diesen standen. Ein Beispiel dieser Sortierung ist in Tabelle 8 darsgetellt.

Originale Rangordnung	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5
	entailment	entailment	not_entailment	entailment	entailment
Neue Rangordnung	Passage 1	Passage 2	Passage 4	Passage 5	Passage 3
	entailment	entailment	entailment	entailment	not_entailment

Tab. 8: Stabiles sortieren der Passagen

Passage 4 wird vor Passage 5 platziert, da sie in der ursprünglichen Reihenfolge bereits höher eingestuft war, obwohl Passage 5 ebenfalls als „entailment“ zur Frage klassifiziert wurde. Diese Anpassung stellt sicher, dass die originale Relevanzbewertung durch das Re-Ranking nicht negativ beeinflusst wird, indem Passagen ohne signifikanten Vorteil in der Klassifizierung einen unverdient höheren Rang erhalten. Diese Art der Sortierung, bei der Elemente mit denselben Sortierkriterien ihre relative Position zueinander beibehalten, wird nach der Definition von Knuth (1973) als stabiles Sortieren bezeichnet.[292]

Demonstration & Evaluation

Das Re-Ranking-Verfahren wurde erfolgreich auf die Wikipedia- und PubMed-Datensätze angewendet, wobei die Klassifikationssysteme keine Anzeichen von Halluzinationen gezeigt haben. Um die Bearbeitungszeit zu verkürzen, wurden nur die Fragen und Passagen-Sets klassifiziert und neu sortiert, bei denen sich die korrekte Passage unter den Top-10 befindet. Dies beeinträchtigt die Evaluationsaussage nicht, da ein Re-Ranking die Position einer nicht in den Top-10 befindlichen Passage nicht verbessern kann und das Ergebnis dieser Fragen und Passage-Sets unverändert bleibt. Zur Bewertung der Abrufungsgenauigkeit werden die Metriken der Top-K Genauigkeit und MRR herangezogen. Die Ergebnisse des Re-Ranking-Verfahrens werden mit den Ergebnissen des ursprünglichen Abrufungssystems ohne dieses Verfahren

[292] Vgl. Knuth 1973, S. 5

verglichen. MRR wird aufgrund seiner Eignung für kleine Mengen betrachteter Passagen ausgewählt[293] und es bewertet wie gut das System die relevante Passage früh in der Rangordnung platziert.[294] Die Top-K Genauigkeit ist relevant, da bei einer korrekten Passage bei kleinerem K weniger Passagen zur Antwortgenerierung benötigt werden, was die Leistung des RAG-Systems durch erhöhte Generierungsgeschwindigkeit signifikant verbessert. Zudem ist die Top-1 Genauigkeit für Deepsearch von Bedeutung, da im Gegensatz zu klassischen Antwortgenerierungsmethoden, bei denen alle Passagen herangezogen werden, Deepsearch die gesamte Seite des Dokuments, auf der sich die Top-1 Passage befindet, als Kontext zur Antwortgenerierung verwendet. Eine Verbesserung der Top-1 Genauigkeit würde demnach erheblich zur Steigerung der Antwortqualität von IBM Deepsearch beitragen.

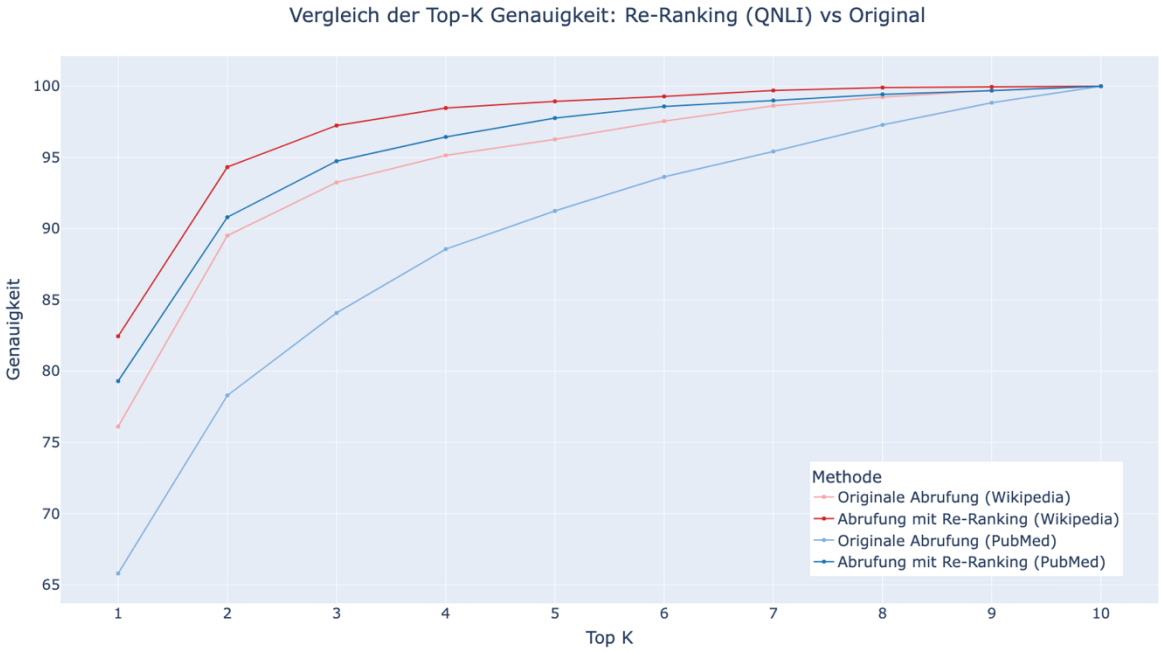


Abb. 11: Top-K Genauigkeit des Re-Rankings mit QNLI-Klassen

Abbildung 11 zeigt die Top-K-Genauigkeit (in Prozent) für K von 1 bis 10 der neuen Abrufungsmethode mit Re-Ranking im Vergleich zur originalen Methode. Es ist erkennbar, dass die Abrufungsmethode von Deepsearch auf dem domänenübergreifenden Wikipedia-Datensatz etwa 10 Prozent höhere Genauigkeit zeigt als auf dem domänenspezifischen PubMed-Datensatz. Auf beiden Datensätzen verbessert das implementierte Re-Ranking die Abrufungsgenauigkeit signifikant. Bei der Top-1-Genauigkeit zeigt sich eine Verbesserung um etwa 6 Prozent auf dem Wikipedia-Datensatz und um etwa 14 Prozent auf dem PubMed-Datensatz. Diese Genauigkeitssteigerung bleibt für die ersten K Passagen bestehen, nimmt jedoch bei höheren K-Werten allmählich ab. Es ist zudem festzustellen, dass das Re-Ranking

[293] Vgl. Shi et al. 2012, S. 140
[294] Die verwendeten Metriken sind in Kapitel 2.3.3 erörtert

die Abrufungsgenauigkeit zu keinem Zeitpunkt insgesamt verschlechtert hat. Die Genauigkeit erreicht bei Top-10 100%, da nur Datenpunkte betrachtet wurden, bei denen sich die korrekte Passage unter den Top-10 befindet. Auch die MRR-Metrik zeigt, dass das Re-Ranking durch die Klassifizierung nach „entailment“ und „not_entailment“ die Relevanzrangordnung der Passagen verbessert hat. Tabelle 9 illustriert die MRR-Werte. Eine detaillierte Darstellung der Werte des Re-Rankings im Vergleich zur originalen Abrufungsmethode befindet sich im Anhang 2/1.

Zusammenfassend lassen die Ergebnisse dieser Iteration darauf schließen, dass ein Re-Ranking nach semantischen Klassen die Genauigkeit eines Abrufungssystems steigern könnte. Diese Ergebnisse werden akzeptiert, und in der nächsten Iteration werden die TREC-Klassen verwendet.

Abrufungsmethode und Datensatz	**MRR**
Originale Abrufung (Wikipedia)	0.8527
Abrufung mit Re-Ranking (Wikipedia)	**0.8991**
Originale Abrufung (PubMed)	0.768
Abrufung mit Re-Ranking (PubMed)	**0.8736**

Tab. 9: MRR des Re-Rankings mit QNLI-Klassen

4.5.2 2. Iteration: TREC-Klassen im Re-Ranking Verfahren

Die in dieser Iteration verwendete Taxonomie wurde in den Kapiteln 2.4 und 4.2.2 detailliert beschrieben.

Entwicklungsphase

Das Re-Ranking der Passagen mit TREC-Klassen folgt demselben Verfahren wie in der vorherigen Iteration, jedoch wird statt der Sortierung nach "entailment" eine Sortierung nach einem binären "matching" Parameter durchgeführt. Dieser Parameter ist erfüllt, wenn die Klassifizierung der Frage mit einer der Klassen der Passage übereinstimmt, wie zum Beispiel, wenn die Frage als "abbreviation" klassifiziert wurde und die Passage die Klassen "abbreviation", "not_entity", "not_human", "location", "numeric", "not_description" enthält. Die Sortierung erfolgt stabil und beschränkt sich auf die Top-10 abgerufenen Passagen.

Demonstration & Evaluation

Das Re-Ranking mit TREC-Klassen wurde erfolgreich implementiert, wobei die Klassifikationssysteme keine Anzeichen von Halluzinationen zeigten. Wie Abbildung 12 zeigt, hat sich die Top-K Genauigkeit durch das Re-Ranking auf den domänenübergreifenden (Wikipedia) und domänenspezifischen (PubMed) Datensätzen signifikant verschlechtert. Bei dem Wikiepdia Datensatz ist die Top-1 Genauigkeit um etwa 22 Prozent gefallen, während sie auf

dem PubMed Datensatz um etwa 20 Prozent gesunken ist. Weiterhin ist festzustellen, dass die Genauigkeitsverschlechterung auf dem PubMed-Datensatz ab Top-6 Passagen gegen Null tendiert, was darauf hindeutet, dass das Re-Ranking ab einer größeren Anzahl von Passagen keinen Einfluss auf die Abrufungsgenauigkeit bei diesem Datensatz hat.

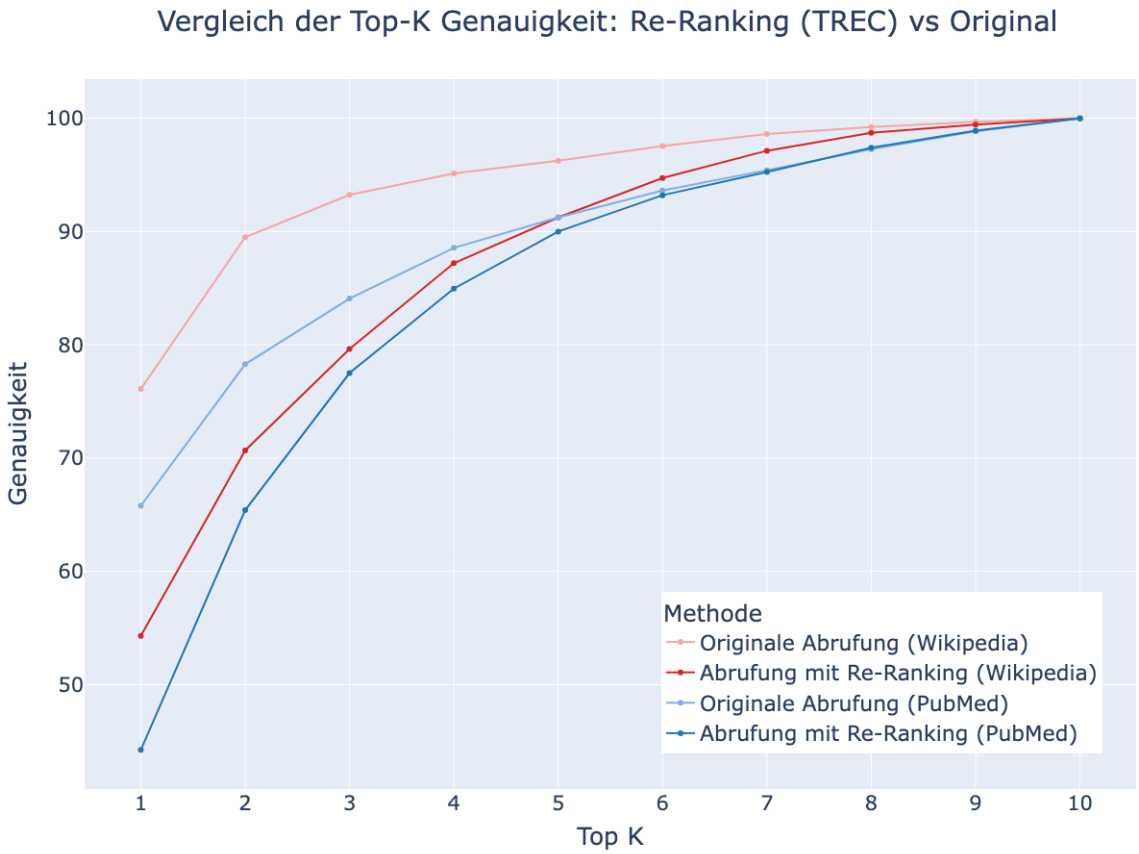


Abb. 12: Top-K Genauigkeit des Re-Rankings mit TREC-Klassen

Tabelle 10 präsentiert die MRR-Werte der verschiedenen Abrufungsmethoden im Vergleich. Auch hieran ist zu erkennen, dass das Re-Ranking mit TREC-Klassen zu insgesamt schlechterer Relevanzordnung der Passagen führt. Eine detaillierte Darstellung der Top-K und MRR-Werte sind in Anhang 2/2 zu finden.

Abrufungsmethode und Datensatz	**MRR**
Originale Abrufung (Wikipedia)	**0.8527**
Abrufung mit Re-Ranking (Wikipedia)	0.6943
Originale Abrufung (PubMed)	**0.768**
Abrufung mit Re-Ranking (PubMed)	0.631

Tab. 10: MRR des Re-Rankings mit TREC-Klassen

Die Ergebnisse der letzten Iteration zeigen zwar, dass eine Steigerung der Abrufungsgenauigkeit durch Re-Ranking nach semantischen Klassen möglich ist, aber der Erfolg entscheidend

von der Angemessenheit der gewählten Klassifizierungsmethoden und Taxonomien abhängig ist. Eine ungeeignete Wahl kann wie in dieser Iteration zu einer signifikanten Verschlechterung führen. Es ist möglich, dass das in Kapitel 4.2.3 entwickelte und ungeprüfte Klassifizierungssystem die Passagen nicht korrekt klassifiziert und somit zu den schlechteren Ergebnissen beiträgt.

4.6 Untersuchung der potentiellen Verbesserung durch eine Integrierte Abrufung

Die nachfolgenden Iterationen sollen, im Einklang mit der zweiten Zielfrage (siehe: Tab. 4) untersuchen, ob eine engere Verknüpfung der ursprünglichen Abrufmethoden mit der semantischen Klassifizierung eine Verbesserung der Abrufungsgenauigkeit ermöglichen kann. Um die engere Integration der semantischen Klassen in die Abrufung zu erreichen, wird in dieser Arbeit der Ansatz verfolgt, die semantische Klassifizierung der Passagen zum skalaren Relevanzwert der Originalpassage zu addieren und somit in die Gesamtrelevanzbewertung einfließen zu lassen. Diese Methodik folgt dem mathematischen Prinzip der hybriden Abrufung, bei der eine Linearkombination der beiden Werte mittels eines Interpolationsparameters oder Gewichts für den hinzuzufügenden Wert vorgenommen wird.[295] Die neue Relevanzbewertung der Passagen wird daher wie folgt definiert:[296]

$$sim_{integrierte\ Abrufung}(q,p) = sim(q^{hyb},p^{hyb}) + \lambda \times \delta(q^{Klasse},p^{Klasse})$$

$$\delta(q^{Klasse},p^{Klasse}) = 1, wenn\ q^{Klasse}\ und\ p^{Klasse}\ zueinander\ passen, sonst\ 0$$

Fragen- und Passagen-Klassen werden bei den TREC-Klassen als zueinander passend betrachtet, wenn die Passagenklassen die entsprechende Frageklasse enthalten (wie Re-Ranking). Im Fall der QNLI-Klassen gelten sie als passend, wenn das Label der Passage „entailment“ ist, was bedeutet, dass die Passage die gestellte Frage beantwortet. Ziel der nachfolgenden Iterationen ist es, das optimale λ zu identifizieren, welches die Genauigkeit der gesamten Relevanzbewertung maximiert.

4.6.1 3. Iteration: QNLI-Klassen in der integrierten Abrufung

Der Ausgangspunkt dieser Iteration sind die skalaren Relevanzwerte, die vom System für jede Passage im Verhältnis zur entsprechenden Frage ausgegeben werden, wie dies bei der Abrufung der Wikipedia- und PubMed-Datensätze durch das Abrufungssystem geschieht. Die integrierte Abrufung wird simuliert, indem die Passagen basierend auf den modifizierten Relevanzwerten neu sortiert werden. Wie in Kapitel 4.5.1 erläutert, werden aufgrund der gegebenen Umstände nur die Top-10 Passagen betrachtet. Das Neusortieren basierend auf der

[295] Vgl. Ji Ma et al. 2021, S. 5
[296] Die Relevanzbewertung durch hybride Abrufung ist in Kapitel 2.3.2 erläutert, die konkrete Implementierung in Kapitel 4.3

neuen Relevanzbewertung führt unter diesen Umständen zu demselben Ergebnis, als ob die Relevanzwerte während der Abrufung in Echtzeit manipuliert würden.

Entwicklungsphase

Da die Relevanzwerte für beide Datensätze im einstelligen bis niedrigen zweistelligen Bereich liegen und lediglich nach einem geeigneten Gewicht für die semantischen Klassen gesucht wird, erfolgt die Suche nach dem optimalen λ in 0.01-Schritten von 0.01 bis 10. Um diese Suche in einem akzeptablen Zeitrahmen umzusetzen, werden die gespeicherten Klassifikationen aus den vorherigen Iterationen verwendet, anstatt die Klassifizierungen erneut durchzuführen. Zudem wird die Gewichtssuche über den gesamten Datensatz vorgenommen, um explorativ zu prüfen, ob ein Gewicht gefunden werden kann, welches die Abrufungsgenauigkeit signifikant verbessert. Sollte für den gesamten Datensatz kein solches Gewicht identifiziert werden können, ist anzunehmen, dass auch eine Aufteilung des Datensatzes in Trainings-, Validierungs- und Testdatensatz keine weiterführenden Ergebnisse liefern wird. Die Optimierung orientiert sich an der MRR-Metrik, da diese im Gegensatz zur Top-K-Genauigkeit einen einzigen vergleichbaren Wert liefert. Zudem impliziert ein höherer MRR-Wert eine verbesserte Top-K-Genauigkeit, da diese Metrik die Position der korrekten Passage innerhalb der Rangordnung bewertet.

Demonstration & Evaluation

Die integrierte Abrufung wurde erfolgreich implementiert und die optimalen Gewichte gefunden. Für den Wikipedia-Datensatz liegt das optimale Gewicht der QNLI-Klassen bei 1,29. Für dieses λ beträgt der MRR-Wert 0,9007.[297] Abbildung 13 präsentiert die Top-K Genauigkeit im Detail.

[297] Visualisierung dieser MRR-Werte in Abhängigkeit von der Gewichtung befindet sich im Anhang 3/1

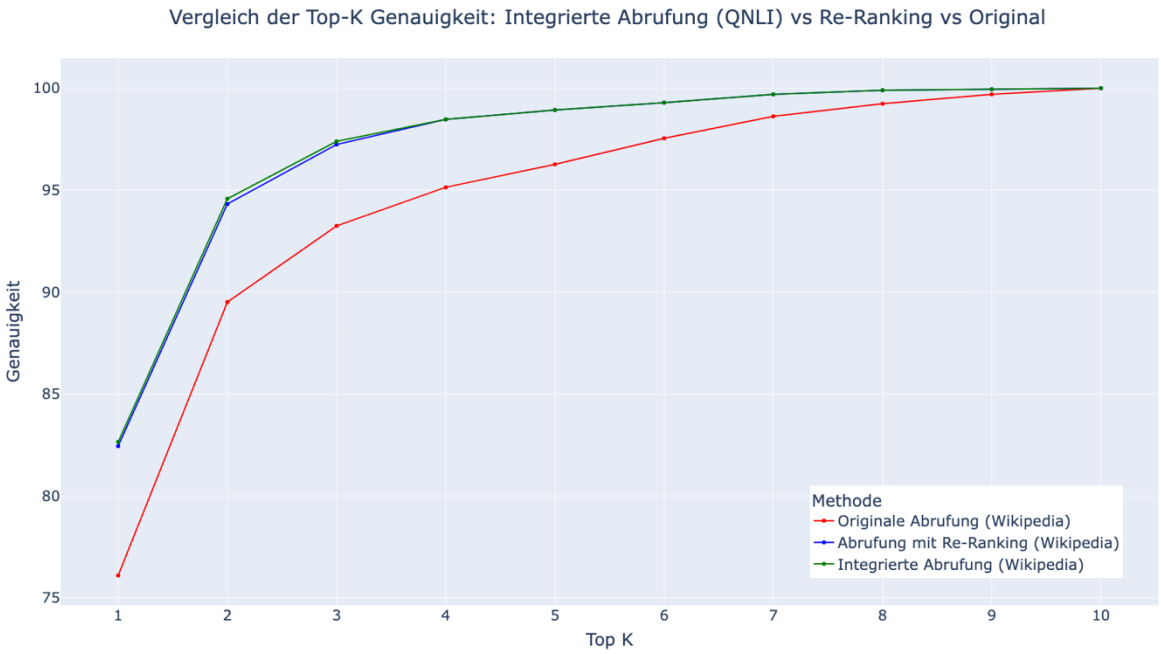


Abb. 13: Top-K Genauigkeit der integrierten Abrufung mit QNLI-Klassen (Wikipedia)

Auf dem Wikipedia-Datensatz zeigt die integrierte Abrufungsmethode nur eine marginale Verbesserung gegenüber dem Re-Ranking mit denselben Klassen. Zudem verschlechterte sich die Abrufungsgenauigkeit mit keinem der getesteten Gewichtungen im Vergleich zur ursprünglichen Methode (ohne Re-Ranking). Der MRR-Wert verbessert sich kontinuierlich ab einem λ -Wert von 0.01 und erreicht bei 1.29 seinen Höhepunkt. Bei weiterer Erhöhung von λ stabilisiert sich der MRR-Wert auf 0,8996 und bleibt bis λ = 10 nahezu unverändert. Dass die Genauigkeit der integrierten Abrufung bei einer geringen Gewichtung der semantischen Klassifizierung nahe der Originalmethode liegt, ist erwartbar, da ein geringes Gewicht bedeutet, dass die semantische Klassifizierung kaum in die Gesamtbewertung einfließt. Dass das optimale Gewicht bei 1,29 liegt, deutet darauf hin, dass die semantische Klassifizierung nur einen kleinen Anteil an der optimalen Gesamtrelevanzbewertung haben sollte. Interessanterweise befindet sich der anteilige Einfluss der semantischen Klassifizierung in ähnlicher Größenordnung wie der Interpolationsparameter (0,1), der in der hybriden Abrufung verwendet wird (siehe Kapitel 4.3).[298]

[298] BM-25 fließt zu 10% in die Hybride Abrufung ein, bei einem Gewicht von 1.29 und einem Relevanzwert von beispielsweise 6 fließt die semantische Klasse zu ca. 22% ein.

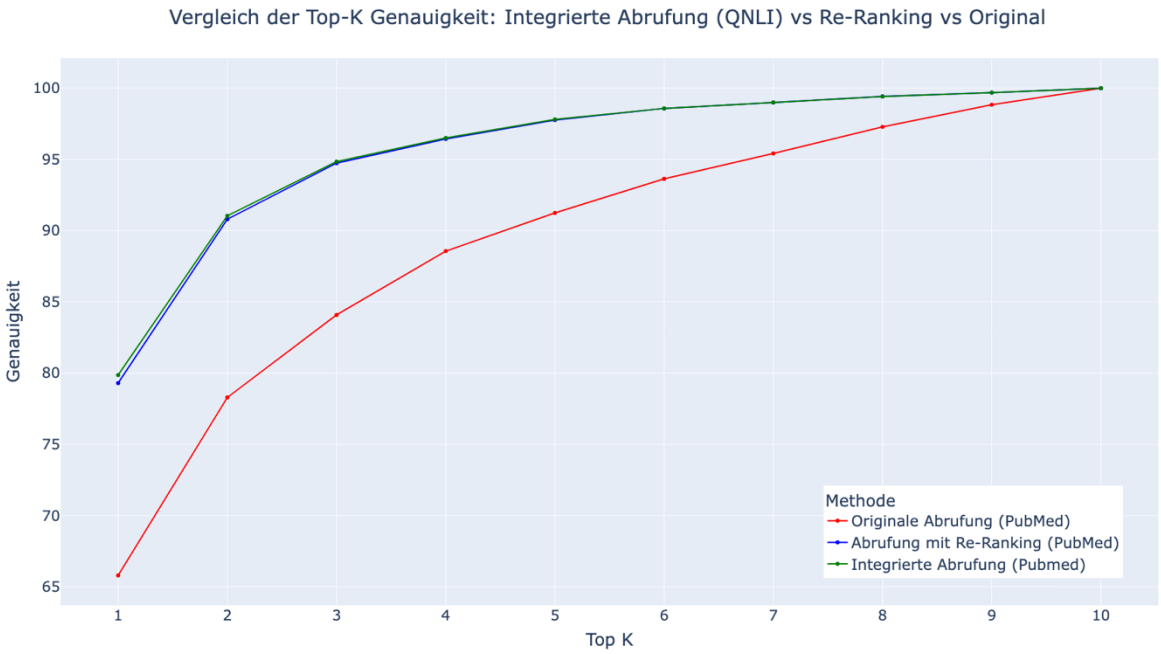


Abb. 14: Top-K Genauigkeit der integrierten Abrufung mit QNLI-Klassen (PubMed)

Für den PubMed-Datensatz liegt die optimale Gewichtung bei λ = 1,48, was zu einem MRR-Wert von 0,877 führt.[299] Ähnlich wie beim Wikipedia-Datensatz zeigt die integrierte Abrufung im Vergleich zur Re-Ranking-Methode nur marginale Verbesserungen in der Top-K Genauigkeit. Die optimale Gewichtung ist vergleichbar zu den Ergebnissen auf dem Wikipedia-Datensatz und die Entwicklung der MRR-Werte in Abhängigkeit von der Gewichtung verläuft ähnlich: Der MRR-Wert steigt kontinuierlich bis zum Optimum an und bleibt dann bis λ = 10 nahezu konstant bei 0,84.

Diese Iteration trägt zur Beantwortung der Zielsetzung bei, indem die Ergebnisse demonstrieren, dass mit einer integrierten Abrufungsmethode basierend auf den QNLI-Klassen eine bessere Abrufungsgenauigkeit als mit der ursprünglichen Methode erreicht werden kann. Allerdings stellt diese Methode nur eine marginale Verbesserung im Vergleich zur Re-Ranking-Methode dar. Zusätzlich bleibt die Frage offen, inwieweit das optimale Gewicht auf bisher nicht berücksichtigte Datensätze generalisierbar ist.

4.6.2 4. Iteration: TREC-Klassen in der integrierten Abrufung

Trotz der Ergebnisse der zweiten Iteration bezüglich des Re-Rankings mit den TREC-Klassen bleibt es eine interessante Untersuchung, ob eine optimale Gewichtung die Abrufungsgenauigkeit gegenüber dem Original verbessern oder zumindest die Verschlechterung reduzieren kann.

[299] Visualisierung dieser MRR-Werte in Abhängigkeit von der Gewichtung befindet sich im Anhang 3/2

Entwicklungsphase

Die Implementierung der integrierten Abrufung mit den TREC-Klassen orientiert sich an den Prinzipien und Konzepten der vorherigen Iteration. Der wesentliche Unterschied besteht darin, dass die Passagenklassen zu den Frageklassen als passend eingestuft werden, wenn die Passagenklasse die Frageklasse beinhaltet. Anstelle des „entailment"-Labels wird ein binäres „matching"-Label verwendet, das gewichtet mit λ zu den ursprünglichen Relevanzwerten hinzuaddiert wird, um die neue Relevanzbewertung zu bilden. Zudem werden die in der zweiten Iteration gespeicherten Klassifizierungsergebnisse verwendet, anstatt die Fragen und Passagen erneut zu klassifizieren.

Demonstration & Evaluation

Die integrierte Abrufung mit den TREC-Klassen wurde erfolgreich umgesetzt, und optimale Gewichtungen wurden ermittelt. Auf dem Wikipedia-Datensatz erreichte die optimale Gewichtung von λ = 0,24 einen MRR-Wert von 0,859,[300] was eine geringe Verbesserung im Vergleich zum ursprünglichen MRR-Wert von 0,853 darstellt. Dieses Ergebnis ist jedoch signifikant da diese minimale Gewichtung vollständig die Verschlechterung durch die Re-Ranking-Methode kompensiert. Auch bei der Top-K Genauigkeit ist erkennbar, dass die integrierte Abrufmethode die ursprüngliche Abrufgenauigkeit bei K = 1 um etwa 1 Prozent verbessert (siehe Abb. 15). Im Gegensatz zur integrierten Abrufung mit QNLI-Klassen beginnt der MRR-Wert für geringe Gewichtungen hoch, erreicht früh bei λ = 0,24 das Maximum und sinkt dann stetig. Für Gewichtungen größer als 6 bleibt der MRR-Wert konstant bei ungefähr 0,7.

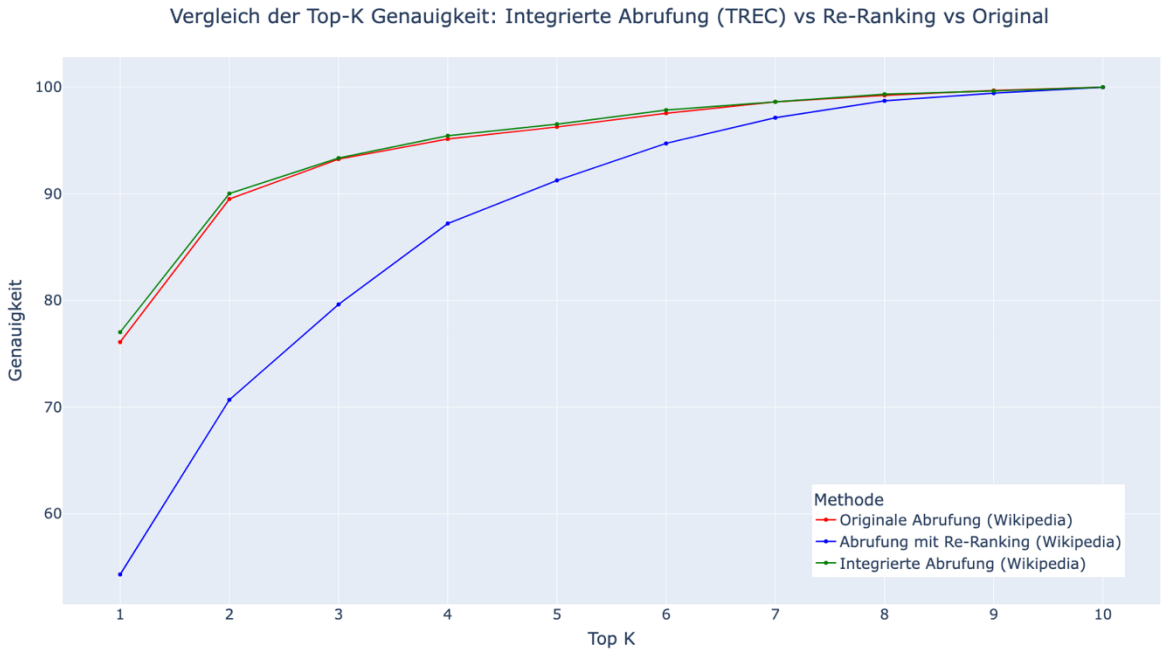


Abb. 15: Top-K Genauigkeit der integrierten Abrufung mit TREC-Klassen (Wikipedia)

[300] Visualisierung dieser MRR-Werte in Abhängigkeit von der Gewichtung befindet sich im Anhang 4/1

Bei der Evaluation auf dem PubMed-Datensatz wurden vergleichbare Resultate festgestellt. Die optimale Gewichtung liegt bei λ = 0,16, was zu einem MRR-Wert von 0,769 führt.[301] Dieser Wert entspricht fast exakt dem MRR der ursprünglichen Abrufmethode.

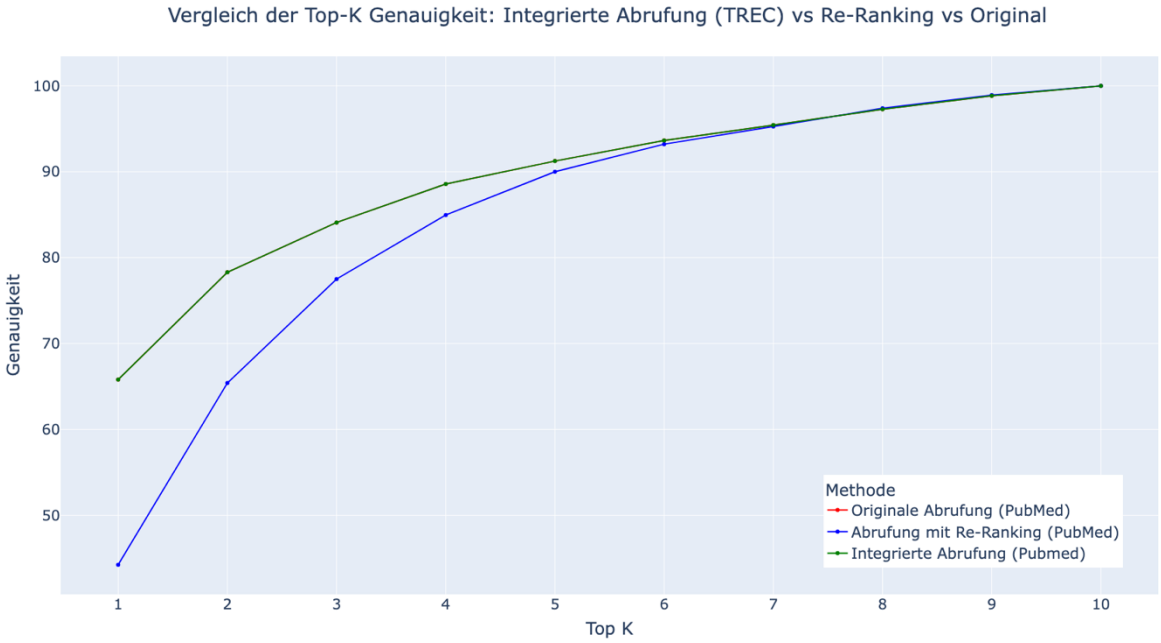


Abb. 16: Top-K Genauigkeit der integrierten Abrufung mit TREC-Klassen (PubMed)

Die annähernd identische Abrufgenauigkeit der neuen Methode im Vergleich zur originären Methode wird ebenso in der Top-K Genauigkeit reflektiert, wobei der Graph der originalen Abrufmethode (rot) unmittelbar unter dem Graphen der integrierten Abrufmethode (grün) verläuft (siehe: Abb. 16).

Die Ergebnisse dieser Iteration deuten darauf hin, dass eine integrierte Abrufmethode signifikant höhere Genauigkeiten erreichen kann als Methoden des Re-Rankings unter Verwendung derselben semantischen Taxonomie. Die, wenn auch geringfügige, Verbesserung der Abrufungsgenauigkeit beim Wikipedia-Datensatz bestätigt die Annahme, dass eine Integration semantischer Klassen generell die Genauigkeit des untersuchten Abrufungssystems verbessern könnte. Insbesondere die Beobachtung, dass das optimale Gewicht gering, jedoch ungleich Null ist, deutet darauf hin, dass eine passendere Taxonomie und ein präziseres Klassifizierungssystem möglicherweise die separate Klassifikation von Fragen und Passagen, wie in der TREC-Taxonomie dargestellt, als effektive Methode für die integrierte Abrufung bestätigen könnte.

[301] Visualisierung dieser MRR-Werte in Abhängigkeit von der Gewichtung befindet sich im Anhang 4/2

5 Ergebnisdiskussion

In Kapitel 4 wurden die in Kapitel 3 definierten Zielsetzungen und Forschungsfragen gemäß dem festgelegten Forschungsdesign untersucht. Zur ersten Frage, ob sich die Abrufungsgenauigkeit durch den Einsatz semantischer Klassifizierung in einem Re-Ranking-Verfahren verbessern lässt, wurde in den ersten beiden Hauptiterationen ein prototypisches Re-Ranking-Verfahren auf den Top-10 abgerufenen Passagen implementiert und anhand des domänenübergreifenden Wikipedia- sowie des domänenspezifischen PubMed-Datensatzes evaluiert. Die Ergebnisse haben gezeigt, dass eine Re-Ranking-Methode mit QNLI-Klassen zu signifikanten Verbesserungen führen kann, während dasselbe stabile Re-Ranking basierend auf TREC-Klassen eine signifikante Verschlechterung der Abrufungsgenauigkeit aufwies. Diese Diskrepanz könnte darauf zurückzuführen sein, dass das Klassifizierungssystem für die QNLI-Klassen mit offiziellen Testdaten evaluiert wurde und eine hohe Klassifizierungsgenauigkeit aufwies. Die Klassifizierung der Passagen nach den TREC-Klassen, ist nich direkt von der Taxonomie vorgesehen und es standen auch keine Testdaten für die Evaluation der Passagenklassifizierung zur Verfügung. Die Annahme einer korrekten Klassifizierung könnte somit irrtümlich sein. Wird jedoch angenommen, dass die Passagenklassifizierung akzeptabel genau erfolgte, legt die Diskrepanz in den Re-Ranking-Ergebnissen nahe, dass es passendere semantische Taxonomien für den vorliegenden Anwendungsfall geben könnte und die TREC-Taxonomie für das Re-Ranking der Abrufungsergebnisse ungeeignet sein könnte.

In den dritten und vierten Hauptiterationen wurde die semantische Klassifizierung in einem integrierten Abrufungsverfahren implementiert und mit denselben Datensätzen evaluiert. Die Abrufungsgenauigkeit der integrierten Abrufung mit QNLI-Klassen war trotz optimaler Gewichtung nur marginal besser als die Re-Ranking-Methode. Angesichts des erhöhten Rechenaufwands bei der Gewichtungsoptimierung lässt dieses Ergebnis vermuten, dass die integrierte Abrufung im Vergleich zum Re-Ranking unter bestimmten Umständen nachteilig sein könnte. Die Ergebnisse der integrierten Abrufung mit TREC-Klassen waren jedoch bemerkenswerter. Aufgrund der schlechten Ergebnisse aus dem Re-Ranking war zu erwarten, dass die optimale Gewichtung bei 0,01 liegen würde, um die TREC-Klassen so gering wie möglich in die Relevanzbewertung einfließen zu lassen. Jedoch lagen die optimalen Gewichte bei 0,24 und 0,16, was in beiden Fällen die Abrufungsgenauigkeit marginal gegenüber dem Original verbesserte und signifikant gegenüber dem Re-Ranking. Dies legt nahe, dass die TREC-Taxonomie in begrenztem Maße Vorteile in der Abrufung bietet und die Methode der separaten Klassifizierung von Fragen und Passagen potenziell bei einer angepassteren Taxonomie die Abrufungsgenauigkeit signifikant verbessern könnte.

Die separate Klassifikation von Fragen und Passagen ist besonders interessant, da die Passagen im Indexierungsschritt des RAG-Systems vorklassifiziert werden können und für die Anwendung der semantischen Klassifikation lediglich die eingegebene Frage zur Laufzeit klassifiziert werden muss. Im Vergleich dazu erfordert die Bestimmung des „entailment“-Labels sowohl die Frage als auch die zugehörige Passage. Dabei muss pro betrachtete Passage dieselbe Frage erneut dem Klassifikationssystem übergeben werden. In dem Kontext der Klassifikation mit vortrainierten Sprachmodellen, wie in dieser Arbeit dargelegt, führt die Reduktion der Klassifizierung auf Fragen zu sowohl laufzeittechnischen als auch wirtschaftlichen Verbesserungen, da eine kürzere Kontextlänge bei der Eingabe zu schnelleren Verarbeitungszeiten führt und die Preise zur Nutzung von Sprachmodellen per API-Zugriff oft von der Anzahl der übergebenen Token abhängen. Eine weitere Überlegung stellt die Frage dar, inwiefern sich die Klassifizierung von Frage und Passage zusammen nach dem „entailment“-Label von dem BERT Cross-Attention-Rescoring unterscheidet. Zwar ist die implementierte Lösung technisch eine Klassifikation, jedoch haben konzeptionell beide Methoden Ähnlichkeiten. Das Sprachmodell darauf zu trainieren, das CLS-Token zwischen dem Cross-Attention der Frage und Passage als Repräsentation der Ähnlichkeit zwischen Frage und Passage zu interpretieren, könnte auch darauf trainiert werden, zu repräsentieren, ob die Passage die Frage beantwortet. Dies wäre eine potenziell untersuchenswerte Hypothese.

6 Kritische Reflexion und Ausblick

Dieses Kapitel bildet den Abschluss dieser Arbeit und dient dazu, die geleistete Arbeit und Ergebnisse kritisch zu reflektieren. Kapitel 6.1 stellt den ursprünglichen Auftrag der Arbeit vor und erörtert, ob die gesetzten Ziele erreicht wurden. In Kapitel 6.2 werden die Ergebnisse und die angewandte Methodik kritisch bewertet. Kapitel 6.3 diskutiert die aus dieser Arbeit resultierenden theoretischen und praktischen Implikationen. Kapitel 6.4 bietet eine zusammenfassende Perspektive auf mögliche zukünftige Forschungsrichtungen.

6.1 Auftrag der Arbeit

Das Ziel dieser Arbeit war es, explorativ zu untersuchen, ob die Abrufleistung eines RAG-Systems, am Beispiel des IBM Deepsearch RAG, durch die Einbindung semantischer Klassifizierungen von Fragen und Passagen verbessert werden kann. Um dieses Ziel zu erreichen, wurden zwei spezifische Forschungsfragen formuliert: Erstens, ob die Genauigkeit des Abrufungssystems durch die Integration semantischer Klassen in einem Re-Ranking-System gesteigert werden kann und zweitens, ob diese Verbesserung durch eine integrierte Abrufung, unter engerer Einbeziehung der semantischen Klassifizierungen, erreicht werden kann. In der Untersuchung wurden etablierte und weit verbreitete Metriken diskutiert und für diesen Anwendungsfall passende ausgewählt. Es wurden anerkannte Klassifizierungstaxonomien erörtert und geeignete Methoden zur Klassifizierung implementiert. Die Ergebnisse dieser Klassifizierungen wurden erfolgreich in den Re-Ranking- und integrierten Abrufungsmethoden verwendet, wobei die integrierte Methode an das etablierte Konzept der hybriden Abrufung angelehnt war. Die Evaluierung der Abrufungsmethoden erfolgte unter kontrollierten Laborbedingungen, war reproduzierbar durchgeführt und dokumentiert. Die Ergebnisse beantworteten die definierten Forschungsfragen und regen zu weiteren Hypothesen an (siehe Kapitel 5). Aufgrund der erfolgreichen Untersuchung und der teilweise erzielten Genauigkeitsverbesserungen wird der zu Beginn dieser Arbeit definierte Auftrag als erfüllt betrachtet. Die erzielten Genauigkeitsverbesserungen und die daraus abgeleiteten Schlussfolgerungen tragen zum übergeordneten Ziel des wissenschaftlichen Diskurses bei, nämlich der Generierung qualitativ hochwertigerer und verlässlicherer Antworten durch RAG-Systeme, und unterstützen die wissenschaftlichen Bemühungen zur weiteren Minderung des Halluzinationsrisikos von generativen Sprachmodellen, welches die Grundproblemstellung für RAG darstellt.

6.2 Kritische Reflexion der Ergebnisse und Methodik

Trotz der ermutigenden Ergebnisse ist eine kritische Reflexion der erzielten Ergebnisse und der angewandten Methodik essentiell, um die Validität der Experimente zu überprüfen und mögliche Limitationen zu identifizieren.

Es ist hervorzuheben, dass die durchgeführten Implementierungen und Evaluierungen grundsätzlich reproduzierbar sind, jedoch bestehen dabei zwei wesentliche Einschränkungen.

Erstens wurden die Klassifizierungsmodelle über eine unternehmensinterne API angesteuert. Obwohl das verwendete Sprachmodell google/flan-ul2 öffentlich zugänglich und die Modellkonfigurationen dokumentiert sind, erfordert eine Reproduktion der Ergebnisse durch externe Parteien eigene Zugänge zu diesem Modell. Zudem ermöglicht die BAM API für interne Nutzung eine hohe API-Bandbreite ohne Limitierungen für die Anzahl der Anfragen, was unter Umständen die Reproduzierbarkeit der Ergebnisse einschränken könnte, besonders bei rechenintensiven Vorgängen wie der Klassifikation der großen Menge an Passagen im Re-Ranking. Der verwendete PubMed-Datensatz darf aufgrund unternehmensinterner Anweisungen nicht öffentlich gemacht werden, was die Nachvollziehbarkeit weiter einschränkt. Der Wikipedia-Datensatz ist hingegen im beigefügten Github-Repository verfügbar.

Die Generalisierbarkeit der Ergebnisse ist ebenfalls kritisch zu betrachten. Obwohl Anstrengungen unternommen wurden, die Evaluierungsdaten repräsentativ für domänenoffene und domänenspezifische Daten zu gestalten, stammen die Datensätze aus unternehmenseigener Produktion. Der genaue Prozess, wie die Dokumente von Wikipedia und PubMed abgefragt und Fragen synthetisch generiert wurden, wurde nicht offengelegt, was die Generalisierbarkeit der Ergebnisse limitiert, da diese Datensätze nicht Teil des wissenschaftlichen Diskurses in diesem Themenbereich sind und ihre Repräsentativität für domänenoffene und spezifische Abrufungsdaten nicht wissenschaftlich verifiziert ist. Dass die Evaluierung auf unternehmenseigenen Daten basierte, bedeutet auch, dass die erzielten Abrufungsgenauigkeiten nicht mit den aktuellen Benchmarks und Werten aus wissenschaftlichen Publikationen vergleichbar sind.

Aufgrund der begrenzten Bearbeitungszeit wurden andere relevante Konzepte nicht getestet, und die neuen Abrufungsmethoden wurden nicht mit alternativen Re-Ranking-Verfahren wie beispielsweise solchen basierend auf Cross-Attention verglichen. Zudem wurde aufgrund der begrenzten Rechenkapazitäten nur die Top-10 der abgerufenen Passagen betrachtet, was zu weniger repräsentativen Ergebnissen führt, da in der Forschung vorwiegend die Abrufungsgenauigkeit über eine größere Anzahl von Passagen gemessen wird.

6.3 Implikationen der Arbeit für Theorie und Praxis

Trotz der in der kritischen Reflexion diskutierten Einschränkungen weisen die Ergebnisse dieser Arbeit Implikationen für den wissenschaftlichen Diskurs und die praktische Anwendung auf.

Eine explizite Auswirkung dieser Ergebnisse auf den wissenschaftlichen Diskurs ist, dass sie zeigen, wie eine zusätzliche semantische Klassifizierung von Fragen und Passagen zu einem hybriden Abrufungssystem signifikante Genauigkeitsverbesserungen liefern kann. Dies könnte den Diskurs über aktuelle Konzepte der Informationsabrufung erweitern und zu bestehenden

Debatten über die Rolle semantischer Klassifikationen beitragen. Wie bereits in früheren Kapiteln erläutert, waren semantische Klassifikationen von Fragen ein häufig verwendetes Konzept in der Informationsabrufung vergangener Generationen. Die vorliegenden Ergebnisse könnten daher als Anstoß dienen, die Eignung semantischer Klassifizierungen für moderne Abrufungssysteme neu zu bewerten.

In der Praxis könnten die Ergebnisse dieser Arbeit vielschichtige Implikationen haben. In der Abteilung, in der diese Forschung durchgeführt wurde, könnte die semantische Klassifizierung weiterentwickelt werden, um möglicherweise Teil des Produktiven RAG-Systems zu werden, falls interne Evaluationen dies als vorteilhaft erachten. Andere Abteilungen innerhalb des Unternehmens, die ebenfalls RAG-Systeme implementieren, könnten erwägen, semantische Klassifikationen in ihre Systeme zu integrieren. Da der Quellcode für die interne BAM-API geschrieben wurde und Daten unternehmensintern frei geteilt werden können, könnten diese Abteilungen die neuen Abrufungsmethoden mit minimalem Entwicklungsaufwand für ihre spezifischen Anwendungen evaluieren und implementieren. Demonstrative Anwendungsfälle, wie die Entwicklung eines Minimum Viable Product (MVP), können nicht nur von der Integration dieser modifizierten Abrufmethode in RAG-Systeme profitieren, sondern auch als Beispiel für die Wertschöpfung durch Watsonx.ai dienen. Watsonx.ai ist hierbei das extern vertriebene Äquivalent zum internen IBM BAM-System.

In der allgemeinen Industriepraxis könnten die Ergebnisse dieser Arbeit als Anregung dienen, die Integration semantischer Klassifizierung mit verschiedenen Methoden und Taxonomien in Abrufungssystemen experimentell zu implementieren und deren Wirksamkeit und Wert für spezifische Anwendungsfälle zu evaluieren.

6.4 Ausblick

Basierend auf der vorliegenden Arbeit lassen sich weiterführende Fragen formulieren und Untersuchungen anstellen. Sollte die in dieser Arbeit präsentierten Ergebnisse akzeptiert und als valide betrachtet werden, könnte erforscht werden, inwiefern die Recheneffizienz dieser Klassifizierung optimiert werden kann. Zum Beispiel könnte die Evaluation effizienterer Klassifikationsmodelle, die keine vortrainierten Sprachmodelle nutzen, erfolgen. Zudem könnte untersucht werden, welches die minimale Anzahl an Parametern oder das kleinste Sprachmodell ist, das eine effektive Klassifikation für eine Taxonomie wie QNLI durchführen kann. Weiterhin könnte das Fine-Tuning kleinerer Sprachmodelle auf eine erhöhte Klassifikationsgenauigkeit, insbesondere bei der Klassifizierung von Passagen wie im Fall der TREC-Taxonomie, hin evaluiert werden. Für Überlegungen zur Implementierung in Produktivsystemen könnte untersucht werden, welche zusätzliche Latenz und Antwortverzögerung durch die Klassifizierung auf das gesamte RAG-System verursacht wird und eine Obergrenze definiert werden, jenseits derer die Methode nicht mehr für dialog- oder chat-basierte RAG-Systeme geeignet wäre.

Ein wesentlicher Aspekt der weiterführenden Forschung wird die kritische Prüfung der hier dargestellten Genauigkeitsverbesserungen sein. Eine grundlegende Untersuchung, die durchgeführt werden muss, um die Validität der Ergebnisse zu bestätigen, ist das Testen auf wissenschaftlich etablierten Benchmark-Abrufungsdaten. Des Weiteren könnte die in Kapitel 5 aufgeworfene Fragestellung untersucht werden, inwiefern sich die Klassifizierung nach „entailment“ von einem Cross-Attention-Rescoring abgrenzen lässt. Es stellt sich auch die Frage, ob ein allgemeines Re-Ranking durch Inferencing implementiert werden könnte, indem ein Sprachmodell durch Prompt-Eingaben evaluiert, welche Passage die Frage am besten beantworten könnte. Weiterhin könnte die Leistung dieser semantischen Klassifizierung im Vergleich zu anderen Abrufungsmethoden evaluiert werden, indem sie mit verschiedenen Abrufungssystemen und Methoden verglichen wird, wie zum Beispiel das Fine-Tuning von Embedding-Modellen, die Verwendung eines Cross-Attention-basierten Re-Rankers oder anderes. Zuletzt könnte auch untersucht werden, wie passendere Taxonomien gestaltet werden könnten, sowohl nach dem Schema eines einzigen Labels für ein Frage- und Passagen-Paar als auch nach dem Schema der getrennten Klassifizierung von Frage und Passage, um die in Kapitel 5 diskutierte höhere Effizienz mit substantiellen Genauigkeitsverbesserungen zu kombinieren.

Anhang

Anhang 1/1: Vollständiger Prompt für das QNLI-Klassifikationsmodell

Classify this question and sentence pair based on its entailment in one of these categories: entailment, not_entailment.

The label 'entailment' when the sentence contains the answer to the question.
The label is labeled 'not_entailment' when the sentence does not contain the answer to the question.

Here are some example questions together with the class label:

Question: Who did NASA recruit by using flawed safety numbers?
Sentence: He concluded that the space shuttle reliability estimate by NASA management was fantastically unrealistic, and he was particularly angered that NASA used these figures to recruit Christa McAuliffe into the Teacher-in-Space program.
Label: entailment

Question: How much solar energy is captured by photosynthesis?
Sentence: Photosynthesis captures approximately 3,000 EJ per year in biomass.
Label: entailment

Question: What does the CAR get help with with regards to communication from ITU-D?
Sentence: In addition, the Central African Republic receives international support on telecommunication related operations from ITU Telecommunication Development Sector (ITU-D) within the International Telecommunication Union to improve infrastructure.
Label: entailment

Question: On Indian Independence Day, kites are flown by citizens which symbolize what concept?
Sentence: Most Delhiites celebrate the day by flying kites, which are considered a symbol of freedom.
Label: entailment

Question: What types of Christianity do Quakers belong to?
Sentence: They include those with evangelical, holiness, liberal, and traditional conservative Quaker understandings of Christianity.
Label: entailment

Question: The Further and Higher Education Act 1992 allows polytechnics to award degrees without what organization's approval?
Sentence: This meant that Polytechnics could confer degrees without the oversight of the national CNAA organization.
Label: entailment

Question: In what square is the theater named after Lee Strasberg located?
Sentence: The Lee Strasberg Theatre and Film Institute is in Union Square, and Tisch School of the Arts is based at New York University, while Central Park SummerStage presents performances of free plays and music in Central Park.
Label: entailment

Question: Who are the famous Venezuelen mandolinist?
Sentence: Today, Venezuelan mandolists include an important group of virtuoso players and ensembles such as Alberto Valderrama, Jesus Rengel, Ricardo Sandoval, Saul Vera, and Cristobal Soto.
Label: entailment

Question: What take place during SWS?
Sentence: System consolidation takes place during slow-wave sleep (SWS).
Label: entailment

Question: Rayon comes from what plant product?
Sentence: Products made from cellulose include rayon and cellophane, wallpaper paste, biobutanol and gun cotton.
Label: entailment

Question: How is it postulated that Mars life might have evolved?
Sentence: Those features can also be observed in algae and cyanobacteria, suggesting that these are adaptations to the conditions prevailing in Antarctica.
Label: not_entailment

Question: What determines how deep a tester will go during regression?
Sentence: They can either be complete, for changes added late in the release or deemed to be risky, or be very shallow, consisting of positive tests on each feature, if the changes are early in the release or deemed to be of low risk.
Label: not_entailment

Question: What was the job title of Ed Policy?
Sentence: Progress on the return stalled, and no announcements were made regarding the future of the league.
Label: not_entailment

Question: What did Ibn Sina receive as payment for helping the emir?
Sentence: Ibn Sina's first appointment was that of physician to the emir, Nuh II, who owed him his recovery from a dangerous illness (997).
Label: not_entailment

Question: How much did Bell et al. try to sell his patent for?
Sentence: By then, the Bell company no longer wanted to sell the patent.
Label: not_entailment

Question: What is Spielberg's most common theme?
Sentence: The notable absence of Elliott's father in E.T., is the most famous example of this theme.
Label: not_entailment

Question: What does the oldest know term for Egypt translate to?
Sentence: The name is of Semitic origin, directly cognate with other Semitic words for Egypt such as the Hebrew מִצְרַיִם (Mitzráyim).
Label: not_entailment

Question: Westminster Abbey was the third highest place of learning after which two places?
Sentence: It was here that the first third of the King James Bible Old Testament and the last half of the New Testament were translated.
Label: not_entailment

Question: What famous encyclopedia contains a Russian back-transliteration of Estonian?
Sentence: It should be noted that Estonian words and names quoted in international publications from Soviet sources are often back-transliterations from the Russian transliteration.
Label: not_entailment

Question: When was The Sound Pattern of English published?

Sentence: An important consequence of the influence SPE had on phonological theory was the downplaying of the syllable and the emphasis on segments.
Label: not_entailment

Question: {row['question']}
Sentence: {row['sentence']}
Label:

Anhang 1/2: Vollständiger Prompt für das TREC-Klassifikationsmodell

Classify this question based on its intent in one of these categories: abbreviation, entity, description, human, location, or numeric.
Focus on the interrogative pronouns.
The result of a query can only consist of a single word. In Example: description
Here are some example questions together with the class label of the question:

Question: What is Mikhail Gorbachev 's middle initial ?
Label: abbreviation

Question: What is the full form of .com ?
Label: abbreviation

Question: What is LMDS ?
Label: abbreviation

Question: What does e.g. stand for ?
Label: abbreviation

Question: When reading classified ads , what does EENTY : other stand for ?
Label: abbreviation

Question: What is a handheld PC ?
Label: description

Question: What does the name Billie mean ?
Label: description

Question: How is Answers.com funded ?
Label: description

Question: What is troilism ?
Label: description

Question: What is the purpose of BIOS ?
Label: description

Question: What is the term for the side of the mountain that faces the prevailing winds ?
Label: entity

Question: What 's played at Wembley Stadium , London , every May ?
Label: entity

Question: What is a fear of color ?
Label: entity

Question: What future movie treat was introduced to American colonists in 1603 by Native Americans ?
Label: entity

Question: What 's the official language of Algeria ?
Label: entity

Question: Who gave Abbie Hoffman his first dose of LSD ?
Label: human

Question: What singer became despondent over the death of Freddie Prinze , quit show business , and then quit the business ?
Label: human

Question: What 19th-century writer had a country estate on the Hudson dubbed Sunnyside ?
Label: human

Question: Name the three races unleashed by the Celestials in Marvel comics .
Label: human

Question: What actor and actress have made the most movies ?
Label: human

Question: Where can I get piano music for the Jamiroquai song Everyday for the midi ?
Label: location

Question: What state is known as the Hawkeye State ?
Label: location

Question: Where does Mother Angelica live ?
Label: location

Question: Where is Erykah Badu originally from ?
Label: location

Question: Where did guinea pigs originate ?
Label: location

Question: How many gallons of water go over Niagra Falls every second ?
Label: numeric

Question: What time of year do most people fly ?
Label: numeric

Question: What is the weight of a teaspoon of matter in a black hole ?
Label: numeric

Question: How many colors are there in a rainbow ?
Label: numeric

Question: When did Mount St. Helen last have a significant eruption ?
Label: numeric

Question: {row['text']}
Label:

Anhang 2/1: Evaluation der Abrufung mit Re-Ranking durch QNLI-Klassen[302]

Top-K	1	2	3	4	5	6	7	8	9	10	MRR
Originale Abrufung (Wikipedia)	76.1	89.51	93.24	95.14	96.26	97.54	98.62	99.23	99.69	100	0.8527
Abrufung mit Re-Ranking (Wikipedia)	**82.45**	**94.32**	**97.24**	**98.46**	**98.93**	**99.28**	**99.69**	**99.9**	**99.95**	100	**0.8991**
Originale Abrufung (PubMed)	65.79	78.29	84.08	88.56	91.24	93.63	95.42	97.28	98.84	100	0.768
Abrufung mit Re-Ranking (PubMed)	**79.3**	**90.8**	**94.74**	**96.43**	**97.76**	**98.58**	**98.99**	**99.43**	**99.68**	100	**0.8736**

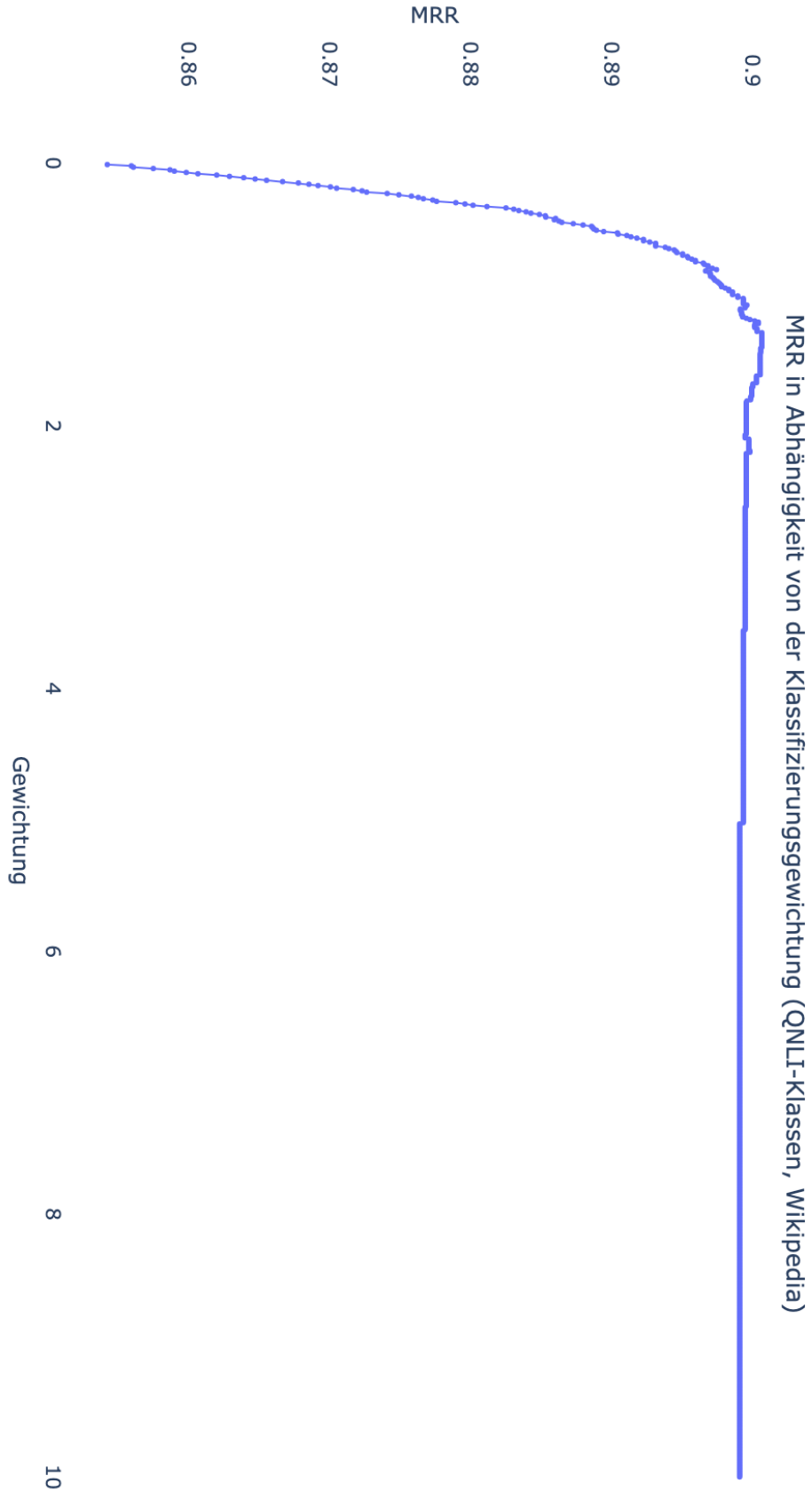
[302] Top-K Genauigkeit in Prozent angegeben, die Werte sind gerundet

Anhang 2/2: Evaluation der Abrufung mit Re-Ranking durch TREC-Klassen[303]

Top-K	1	2	3	4	5	6	7	8	9	10	MRR
Originale Abrufung (Wikipedia)	76.1	89.51	93.24	95.14	96.26	97.54	98.62	99.23	99.69	100	0.8527
Abrufung mit Re-Ranking (Wikipedia)	**54.3**	**70.68**	**79.63**	**87.21**	**91.25**	**94.73**	**97.13**	**98.72**	**99.44**	100	**0.6943**
Originale Abrufung (PubMed)	65.79	78.29	84.08	88.56	91.24	93.63	95.42	97.28	98.84	100	0.768
Abrufung mit Re-Ranking (PubMed)	**44.24**	**65.4**	**77.5**	**84.96**	**90**	**93.21**	**95.26**	**97.4**	**98.91**	100	**0.631**

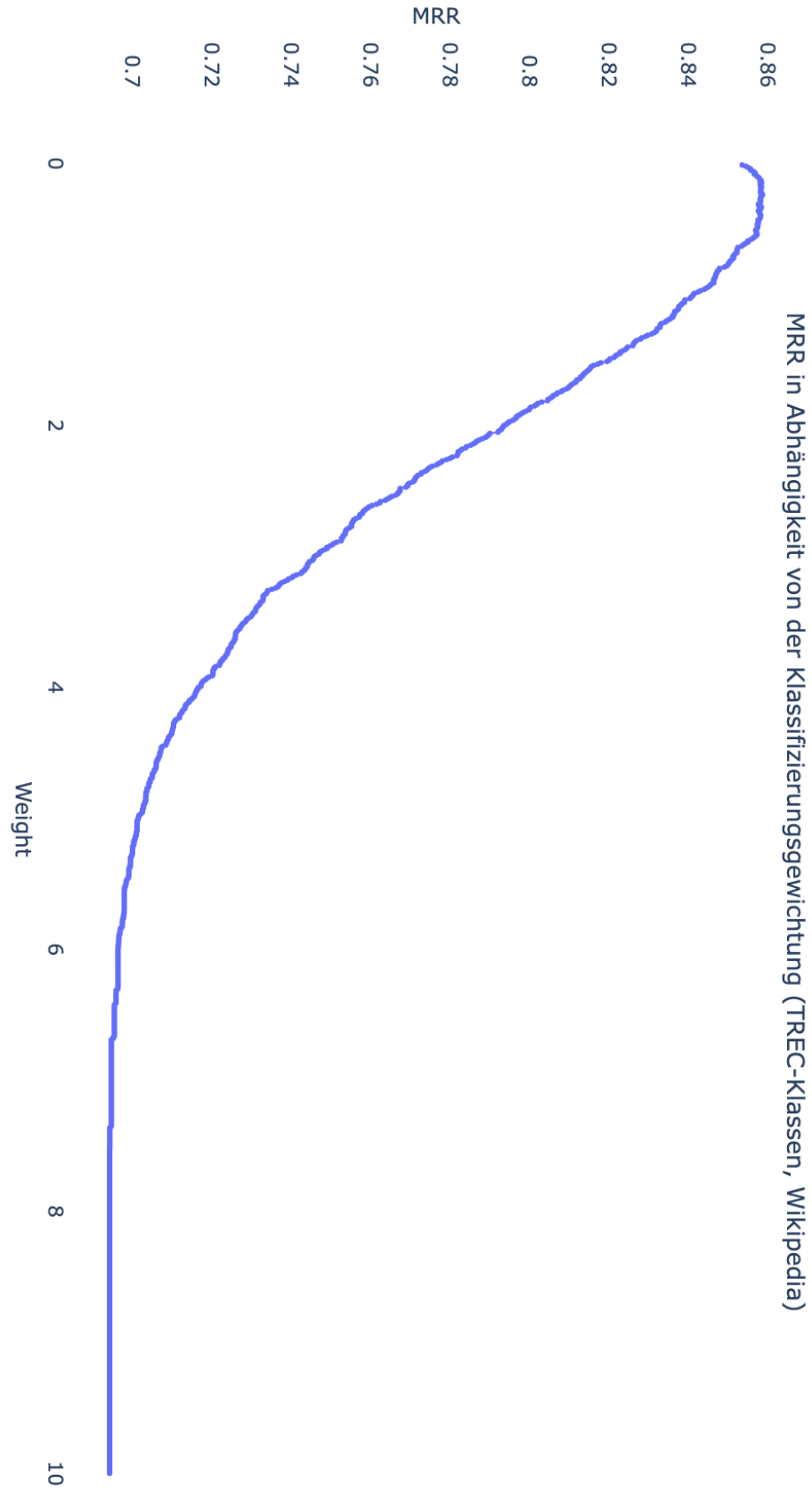
[303] Top-K Genauigkeit in Prozent angegeben, die Werte sind gerundet

Anhang 3/1: Gewichtsoptimierung (QNLI, Wikipedia)

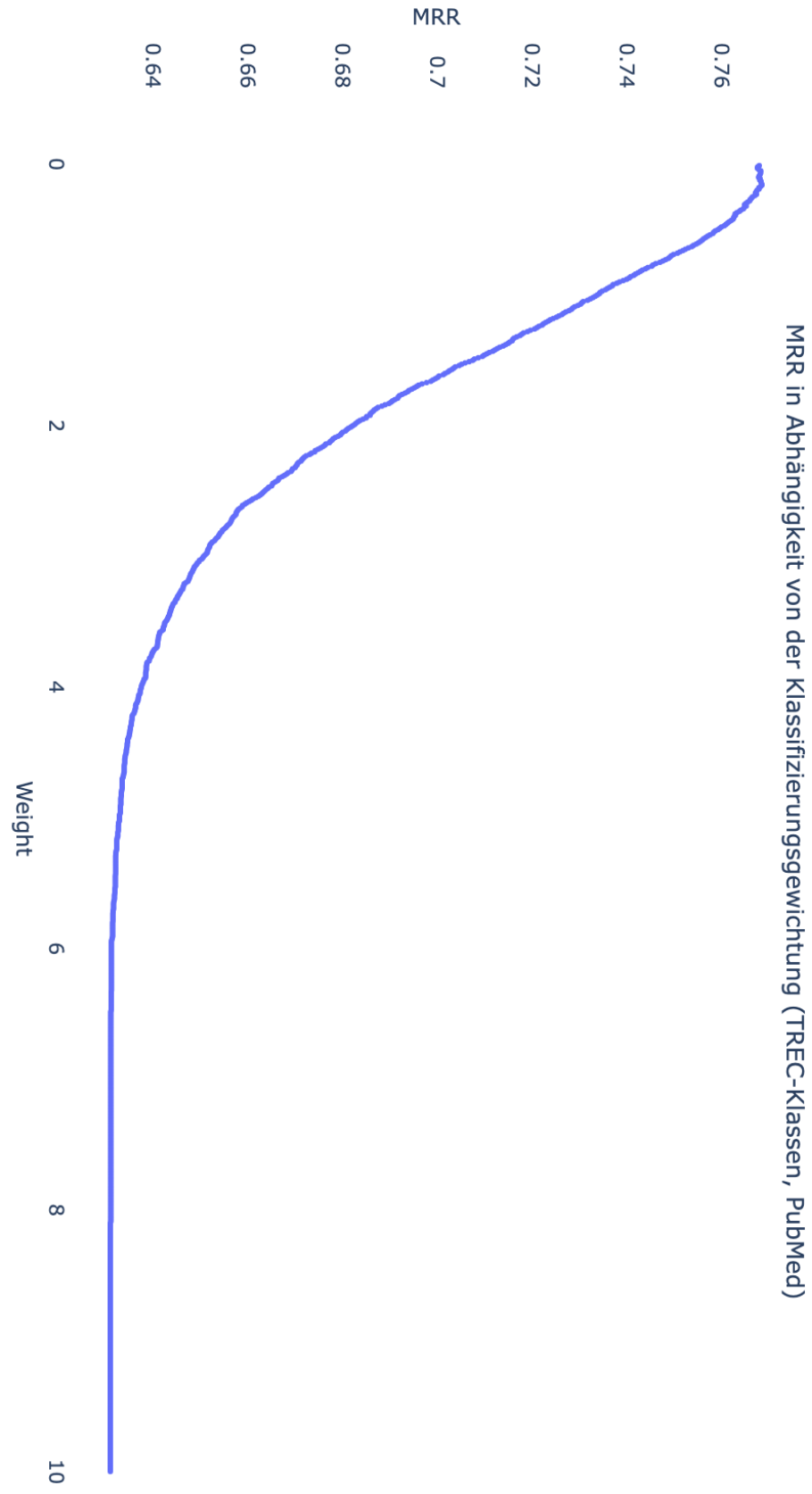


Anhang 3/2: Gewichtsoptimierung (QNLI, PubMed)

Anhang 4/1: Gewichtungsoptimierung (TREC, Wikipedia)



Anhang 4/2: Gewichtungsoptimierung (TREC, PubMed)



Literaturverzeichnis

Alberti, Chris et al. 2019. »Synthetic QA Corpora Generation with Roundtrip Consistency«, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, hrsg. v. Korhonen, Anna; Traum, David; Màrquez, Lluís, S. 6168–6173. Florence, Italy: Association for Computational Linguistics.

Allam, Ali; Haggag, Mohamed 2012. »The Question Answering Systems: A Survey«, in *International Journal of Research and Reviews in Information Sciences* 2, S. 211–221.

Alquaary, Abdalrhman; Çelebi, Numan 2023. »FlexiGPT: Engaging with Documents«, in *Cognitive Models and Artificial Intelligence Conference Proceedings*, S. 87–91. SETSCI.

Arabzadeh, Negar; Yan, Xinyi; Clarke, Charles L. A. 2021. »Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection«, in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, S. 2862–2866. New York, NY, USA: Association for Computing Machinery.

Arora, Monika; Kanjilal, Uma; Varshney, Dinesh 2016. »Evaluation of information retrieval: precision and recall«, in *International Journal of Indian Culture and Business Management* 12, 2, S. 224.

Béchard, Patrice; Ayala, Orlando Marquez 2024. *Reducing hallucination in structured outputs via Retrieval-Augmented Generation*. arXiv.

Blattmann, Andreas et al. 2022. »Retrieval-Augmented Diffusion Models«, hrsg. v. Koyejo, Sanmi et al. Red Hook, NY: Ludwig-Maximilians-Universität München.

Breuel, T. 2003. »Information Extraction from HTML Documents by Structural Matching«, in *Second International Workshop on Web Document Analysis. International Workshop on Web Document Analysis (WDA-2003), located at ICDAR 2003*. Edinburgh.

Brown, Tom et al. 2020. »Language Models are Few-Shot Learners«, in *Advances in Neural Information Processing Systems*, S. 1877–1901. Curran Associates, Inc.

Bruch, Sebastian; Gai, Siyu; Ingber, Amir 2024. »An Analysis of Fusion Functions for Hybrid Retrieval«, in *ACM Transactions on Information Systems* 42, 1, S. 1–35.

Buckley, Chris; Voorhees, Ellen M. 2000. »Evaluating evaluation measure stability«, in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, S. 33–40. New York, NY, USA: Association for Computing Machinery.

Carrara, Fabio et al. 2022. »Approximate Nearest Neighbor Search on Standard Search Engines«, in *Similarity Search and Applications*, hrsg. v. Skopal, Tomáš et al., S. 214–221. Cham: Springer International Publishing.

Chang, Wei-Cheng et al. 2020b. »Pre-training Tasks for Embedding-based Large-scale Retrieval«.

Chen, Danqi et al. 2017b. »Reading Wikipedia to Answer Open-Domain Questions«, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, hrsg. v. Barzilay, Regina; Kan, Min-Yen, S. 1870–1879. Vancouver, Canada: Association for Computational Linguistics.

Chen, Tong et al. 2023. »Dense X Retrieval: What Retrieval Granularity Should We Use?«

Chen, Xilun; Lakhotia, Kushal; Oguz, Barlas; et al. 2022. »Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One?«, in *Findings of the Association for Computational Linguistics: EMNLP 2022*, hrsg. v. Goldberg, Yoav; Kozareva, Zornitsa; Zhang, Yue, S. 250–262. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chen, Yuyan et al. 2023. »Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models«, in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, S. 245–255. New York, NY, USA: Association for Computing Machinery.

Chiu, Chih-Yi; Prayoonwong, Amorntip; Liao, Yin-Chih 2020. »Learning to Index for Nearest Neighbor Search«, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 8, S. 1942–1956.

Christian, Hans; Agus, Mikhael Pramodana; Suhartono, Derwin 2016. »Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)«, in *ComTech: Computer, Mathematics and Engineering Applications* 7, 4, S. 285.

Cleverley, Paul H; Burnett, Simon 2019. »Enterprise search: A state of the art«, in *Business Information Review* 36, 2, S. 60–69.

Cormack, Gordon V.; Lynam, Thomas R. 2006. »Statistical precision of information retrieval evaluation«, in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, S. 533–540. Seattle Washington USA: ACM.

Cummins, Ronan 2013. »A Standard Document Score for Information Retrieval«, in *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, S. 113–116. Copenhagen Denmark: ACM.

Devlin, Jacob et al. 2019. »BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding«, in *Proceedings of the 2019 Conference of the North*, S. 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Du, Jingfei et al. 2020b. »General Purpose Text Embeddings from Pre-trained Language Models for Scalable Inference«, in *Findings of the Association for Computational Linguistics: EMNLP 2020*, hrsg. v. Cohn, Trevor; He, Yulan; Liu, Yang, S. 3018–3030. Online: Association for Computational Linguistics.

Fu, Xuan et al. 2023. »SS-BERT: A Semantic Information Selecting Approach for Open-Domain Question Answering«, in *Electronics* 12, 7, S. 1692.

Galar, Mikel et al. 2011. »An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes«, in *Pattern Recognition* 44, 8, S. 1761–1776.

Glass, Michael et al. 2022. »Re2G: Retrieve, Rerank, Generate«, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, S. 2701–2715. Seattle, United States: Association for Computational Linguistics.

Guo, Mandy et al. 2021. »MultiReQA: A Cross-Domain Evaluation forRetrieval Question Answering Models«, in *Proceedings of the Second Workshop on Domain Adaptation for NLP*, hrsg. v. Ben-David, Eyal et al., S. 94–104. Kyiv, Ukraine: Association for Computational Linguistics.

Guo, Xu; Yu, Han 2022. *On the Domain Adaptation and Generalization of Pretrained Language Models: A Survey*. arXiv.

Guu, Kelvin et al. 2020b. »REALM: retrieval-augmented language model pre-training«, in *Proceedings of the 37th International Conference on Machine Learning*, S. 3929–3938. JMLR.org.

Harman, Donna K. 1993. *The First Text REtrieval Conference (TREC-1)*. U.S. Department of Commerce, National Institute of Standards and Technology.

Holzweißig, Kai 2017. *Wissenschaftliches Arbeiten*. Leanpub.

Ijesunor Akhigbe, Bernard; Samuel Afolabi, Babajide; Rotimi Adagunodo, Emmanuel 2011. »Assessment of Measures for Information Retrieval System Evaluation: A Usercentered Approach«, in *International Journal of Computer Applications* 25, 7, S. 6–12.

Kamps, Jaap et al. Hrsg. 2023. *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part II*. Cham: Springer Nature Switzerland.

Karpukhin, Vladimir; Oguz, Barlas; et al. 2020. »Dense Passage Retrieval for Open-Domain Question Answering«, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, hrsg. v. Webber, Bonnie et al., S. 6769–6781. Online: Association for Computational Linguistics.

Kathare, Nikita; O. Vinati, Reddy; Prabhu, Vishalakshi 2022. »A Comprehensive Study of Elastic Search«, in *Journal of Research in Science and Engineering* 4, 11.

Khramtsova, Ekaterina et al. 2023. »Selecting which Dense Retriever to use for Zero-Shot Search«, in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, S. 223–233. New York, NY, USA: Association for Computing Machinery.

Kishida, Kazuaki 2005. »Property of average precision and its generalization: an examination of evaluation indicator for information retrieval«.

Knuth, Donald Ervin 1973. *The art of computer programming*. Reading, Mass: Addison-Wesley Pub. Co.

Kolomiyets, Oleksandr; Moens, Marie-Francine 2011. »A survey on question answering technology from an information retrieval perspective«, in *Information Sciences* 181, 24, S. 5412–5434.

Korra, R. et al. 2011. »Performance evaluation of Multilingual Information Retrieval (MLIR) system over Information Retrieval (IR) system«, in *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, S. 722–727. Chennai, Tamil Nadu: IEEE.

Lee, Jinhyuk; Wettig, Alexander; Chen, Danqi 2021b. »Phrase Retrieval Learns Passage Retrieval, Too«, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, hrsg. v. Moens, Marie-Francine et al., S. 3661–3672. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Lee, Kenton; Chang, Ming-Wei; Toutanova, Kristina 2019b. »Latent Retrieval for Weakly Supervised Open Domain Question Answering«, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* , S. 6086–6096.

Lewis, Patrick et al. 2020. »Retrieval-augmented generation for knowledge-intensive NLP tasks«, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, S. 9459–9474. Red Hook, NY, USA: Curran Associates Inc.

Li, Huayang et al. 2022b. »A Survey on Retrieval-Augmented Text Generation«, in *ArXiv* .

Li, Jiaqi et al. 2023. *LooGLE: Can Long-Context Language Models Understand Long Contexts?*. arXiv.

Li, Minghan; Gaussier, Eric 2022. »BERT-based Dense Intra-ranking and Contextualized Late Interaction via Multi-task Learning for Long Document Retrieval«, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* , S. 2347–2352.

Li, Wen et al. 2020. »Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement«, in *IEEE Transactions on Knowledge and Data Engineering* 32, 8, S. 1475–1488.

Li, Xin; Roth, Dan 2002. »Learning question classifiers«, in *Proceedings of the 19th international conference on Computational linguistics* -, S. 1–7. Taipei, Taiwan: Association for Computational Linguistics.

Livathinos, Nikolaos et al. 2021. »Robust PDF Document Conversion using Recurrent Neural Networks«, in *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17, S. 15137–15145.

Lv, Yuanhua; Zhai, ChengXiang 2011. »Lower-bounding term frequency normalization«, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, S. 7–16. Glasgow Scotland, UK: ACM.

Ma, Ji et al. 2021. »Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation«, in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, hrsg. v. Merlo, Paola; Tiedemann, Jorg; Tsarfaty, Reut, S. 1075–1088. Online: Association for Computational Linguistics.

Ma, Xinbei et al. 2023b. »Query Rewriting in Retrieval-Augmented Large Language Models«.

Ma, Xueguang et al. 2021. *A Replication Study of Dense Passage Retriever*. arXiv.

Ma, Xueguang et al. 2022. »Another Look at DPR: Reproduction of Training and Replication of Retrieval«, in *Advances in Information Retrieval*, hrsg. v. Hagen, Matthias et al., S. 613–626. Cham: Springer International Publishing.

MacAvaney, Sean et al. 2019. »CEDR: Contextualized Embeddings for Document Ranking«, in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 1101–1104.

Macdonald, Craig; Tonellotto, Nicola 2021. »On Approximate Nearest Neighbour Selection for Multi-Stage Dense Retrieval«, in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, S. 3318–3322. Virtual Event Queensland Australia: ACM.

Malkov, Yu A.; Yashunin, D. A. 2020. »Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs«, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 4, S. 824–836.

Mallen, Alex et al. 2023. »When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories«, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, hrsg. v. Rogers, Anna; Boyd-Graber, Jordan; Okazaki, Naoaki, S. 9802–9822. Toronto, Canada: Association for Computational Linguistics.

Mao, Yuning et al. 2021b. »Reader-Guided Passage Reranking for Open-Domain Question Answering«, in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* , S. 344–350.

Mao, Yuning et al. 2021c. »Generation-Augmented Retrieval for Open-Domain Question Answering«, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, hrsg. v. Zong, Chengqing et al., S. 4089–4100. Online: Association for Computational Linguistics.

McSherry, Frank; Najork, Marc 2008. »Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores«, in *Advances in Information Retrieval*, hrsg. v. Macdonald, Craig et al., S. 414–421. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mishra, Amit; Jain, Sanjay Kumar 2016. »A survey on question answering systems with classification«, in *Journal of King Saud University - Computer and Information Sciences* 28, 3, S. 345–361.

Mitra, Bhaskar; Craswell, Nick 2017. *Neural Models for Information Retrieval*. arXiv.

Modarressi, Ali et al. 2023b. »RET-LLM: Towards a General Read-Write Memory for Large Language Models«.

Moldovan, D. et al. 1999. »LASSO: A Tool for Surfing the Answer Net«.

Mosbach, Marius et al. 2023. »Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation«, in *Findings of the Association for Computational Linguistics: ACL 2023*, hrsg. v. Rogers, Anna; Boyd-Graber, Jordan; Okazaki, Naoaki, S. 12284–12314. Toronto, Canada: Association for Computational Linguistics.

Nair, Suraj et al. 2022. »Transfer Learning Approaches for Building Cross-Language Dense Retrieval Models«, in *Advances in Information Retrieval*, hrsg. v. Hagen, Matthias et al., S. 382–396. Cham: Springer International Publishing.

Nassar, Ahmed et al. 2022. »TableFormer: Table Structure Understanding with Transformers«, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 4604–4613.

Neves, Mariana 2015. »HPI Question Answering System in the BioASQ 2015 Challenge«.

Nogueira, Rodrigo; Cho, Kyunghyun 2020. *Passage Re-ranking with BERT*. arXiv.

Oguz, Barlas et al. 2022. »Domain-matched Pre-training Tasks for Dense Retrieval«, in *Findings of the Association for Computational Linguistics: NAACL 2022*, hrsg. v. Carpuat, Marine; de Marneffe, Marie-Catherine; Meza Ruiz, Ivan Vladimir, S. 1524–1534. Seattle, United States: Association for Computational Linguistics.

Otero, David; Parapar, Javier; Barreiro, Álvaro 2023. »Relevance feedback for building pooled test collections«, in *Journal of Information Science* .

Paramasivam, Aarthi; Nirmala, S. Jaya 2022. »A survey on textual entailment based question answering«, in *Journal of King Saud University - Computer and Information Sciences* 34, 10, S. 9644–9653.

Peffers, Ken et al. 2007. »A Design Science Research Methodology for Information Systems Research«, in *Journal of Management Information Systems* 24, 3, S. 45–77.

Peng, Hao et al. 2019. »Text Generation with Exemplar-based Adaptive Decoding«, in *Proceedings of the 2019 Conference of the North*, S. 2555–2565. Minneapolis, Minnesota: Association for Computational Linguistics.

Prakash, Prafull; Killingback, Julian; Zamani, Hamed 2021. »Learning Robust Dense Retrieval Models from Incomplete Relevance Labels«, in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 1728–1732. Virtual Event Canada: ACM.

Qaiser, Shahzad; Ali, Ramsha 2018. »Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents«, in *International Journal of Computer Applications* 181, 1, S. 25–29.

Radev, Dragomir et al. 2002. »Probabilistic question answering on the web«, in *Proceedings of the 11th international conference on World Wide Web*, S. 408–419. New York, NY, USA: Association for Computing Machinery.

Radlinski, Filip; Craswell, Nick 2010. »Comparing the sensitivity of information retrieval metrics«, in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, S. 667–674. Geneva Switzerland: ACM.

Ramos, J. E. 2003. »Using TF-IDF to Determine Word Relevance in Document Queries«.

Ramponi, Alan; Plank, Barbara 2020. »Neural Unsupervised Domain Adaptation in NLP—A Survey«, in *Proceedings of the 28th International Conference on Computational Linguistics*, hrsg. v. Scott, Donia; Bel, Nuria; Zong, Chengqing, S. 6838–6855. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Rawte, Vipula; Sheth, Amit; Das, Amitava 2023. »A Survey of Hallucination in Large Foundation Models«.

Reichman, Benjamin; Heck, Larry 2024. *Retrieval-Augmented Generation: Is Dense Passage Retrieval Retrieving?*. arXiv.

Ren, Ruiyang et al. 2021. »RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking«, in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, hrsg. v. Moens, Marie-Francine et al., S. 2825–2835. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Reynolds, Laria; McDonell, Kyle 2021. »Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm«, in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, S. 1–7. Yokohama Japan: ACM.

Rifkin, Ryan; Klautau, Aldebaro 2004. »In Defense of One-Vs-All Classification«, in *Journal of Machine Learning Research* 5, S. 101–141.

Roberts, Adam; Raffel, Colin; Shazeer, Noam 2020b. »How Much Knowledge Can You Pack Into the Parameters of a Language Model?«, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, hrsg. v. Webber, Bonnie et al., S. 5418–5426. Online: Association for Computational Linguistics.

Robertson, S. et al. 1995. »Okapi at TREC-4«.

Robertson, Stephen; Zaragoza, Hugo 2009. »The Probabilistic Relevance Framework: BM25 and Beyond«, in *Foundations and Trends® in Information Retrieval* 3, 4, S. 333–389.

Robertson, Stephen; Zaragoza, Hugo; Taylor, Michael 2004. »Simple BM25 extension to multiple weighted fields«, in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, S. 42–49. Washington D.C. USA: ACM.

Rosa, Guilherme Moraes et al. 2021. *Yes, BM25 is a Strong Baseline for Legal Case Retrieval*. arXiv.

Sachan, Devendra Singh et al. 2023. »Questions Are All You Need to Train a Dense Passage Retriever«, in *Transactions of the Association for Computational Linguistics* 11, S. 600–616.

Sangodiah, Anbuselvan; Muniandy, Manoranjitham; Heng, L.E. 2015. »Question classification using statistical approach: A complete review«, in *Journal of Theoretical and Applied Information Technology* 71, S. 386–395.

Saracevic, Tefko 1995. »Evaluation of evaluation in information retrieval«, in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '95*, S. 138–146. Seattle, Washington, United States: ACM Press.

Sarrouti, Mourad; Ouatik El Alaoui, Said 2017. »A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering«, in *Journal of Biomedical Informatics* 68, S. 96–103.

Seo, Minjoon; Lee, Jinhyuk; Kwiatkowski, Tom; Parikh, Ankur; et al. 2019. »Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index«, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, hrsg. v. Korhonen, Anna; Traum, David; Màrquez, Lluís, S. 4430–4441. Florence, Italy: Association for Computational Linguistics.

Shi, Yue et al. 2012. »CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering«, in *Proceedings of the sixth ACM conference on Recommender systems*, S. 139–146. Dublin Ireland: ACM.

Siriwardhana, Shamane et al. 2023. »Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering«, in *Transactions of the Association for Computational Linguistics* 11, S. 1–17.

Soboroff, Ian 2022. »Overview of TREC 2021«, in *NIST* .

van Solingen (Revision), Rini et al. 2002. »Goal Question Metric (GQM) Approach«, in *Encyclopedia of Software Engineering*. John Wiley & Sons, Ltd.

Soudani, Heydar; Kanoulas, Evangelos; Hasibi, Faegheh 2024. *Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge*. arXiv.

Sujatha, Pothula; Dhavachelvan, P. 2011. »Precision at K in Multilingual Information Retrieval«, in *International Journal of Computer Applications* 24, 9, S. 40–43.

Sun, Xiaofei et al. 2023. »Text Classification via Large Language Models«, in *Findings of the Association for Computational Linguistics: EMNLP 2023*, S. 8990–9005. Association for Computational Linguistics.

Talman, Aarne et al. 2022. »How Does Data Corruption Affect Natural Language Understanding Models? A Study on GLUE datasets«, in *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, hrsg. v. Nastase, Vivi et al., S. 226–233. Seattle, Washington: Association for Computational Linguistics.

Tayyar Madabushi, Harish; Lee, Mark; Barnden, John 2018. »Integrating Question Classification and Deep Learning for improved Answer Selection«, in *Proceedings of the 27th International Conference on Computational Linguistics*, hrsg. v. Bender, Emily M.; Derczynski, Leon; Isabelle, Pierre, S. 3283–3294. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Tschopp, Dominique; Diggavi, Suhas 2009. *Approximate Nearest Neighbor Search through Comparisons*. arXiv.

Unlu, Ozan et al. 2024. *Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening*. Health Informatics.

Valizadegan, Hamed et al. 2009. »Learning to Rank by Optimizing NDCG Measure«, in *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Voorhees, Ellen M. et al. 2020. »Overview of the TREC 2019 Deep Learning Track«, in *NIST* .

Wang, Alex et al. 2019. »GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding«, in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, S. 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, Boxin et al. 2021. »Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models«, in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1.

Wang, Shuai; Zhuang, Shengyao; Zuccon, Guido 2021. »BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval«, in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, S. 317–324. Virtual Event Canada: ACM.

Wang, Weishi; Wang, Yue; Joty, Shafiq; Hoi, Steven C.H. 2023. »RAP-Gen: Retrieval-Augmented Patch Generation with CodeT5 for Automatic Program Repair«, in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, S. 146–158. New York, NY, USA: Association for Computing Machinery.

Wang, Yubo; Ma, Xueguang; Chen, Wenhu 2024. *Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering*. arXiv.

Wang, Zhiguo et al. 2019b. »Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering«, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, hrsg. v. Inui, Kentaro et al., S. 5878–5882. Hong Kong, China: Association for Computational Linguistics.

Webster, Jane; Watson, Richard 2002. »Analyzing the Past to Prepare for the Future: Writing a Literature Review«, in *MIS Quarterly* 26.

Whissell, John S.; Clarke, Charles L. A. 2011. »Improving document clustering using Okapi BM25 feature weighting«, in *Information Retrieval* 14, 5, S. 466–487.

Widodo, Sulisetyo Puji 2023. »Comparative Analysis of Retriever and Reader for Open Domain Questions Answering on BPS Knowledge in Indonesian«, in *Proceedings of The International Conference on Data Science and Official Statistics* 2023, S. 337–343.

Wilde, Thomas; Hess, Thomas 2007. »Forschungsmethoden der Wirtschaftsinformatik«, in *WIRTSCHAFTSINFORMATIK* 49, 4, S. 280–287.

Wong, Wilson Kia Onn 2024. »The sudden disruptive rise of generative artificial intelligence? An evaluation of their impact on higher education and the global workplace«, in *Journal of Open Innovation: Technology, Market, and Complexity* 10, 2, S. 100278.

Wu, Zhijing et al. 2019. »Investigating Passage-level Relevance and Its Role in Document-level Relevance Judgment«, in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 605–614. Paris France: ACM.

Xian, Yongqin et al. 2019. »Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly«, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 09, S. 2251–2265.

Xiong, Lee; Xiong, Chenyan; Li, Ye; Tang, Kwok-Fung; Liu, Jialin; Bennett, Paul N.; et al. 2020. »Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval«.

Xu, Benfeng et al. 2023. »Retrieval-Augmented Domain Adaptation of Language Models«, in *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, S. 54–64. Toronto, Canada: Association for Computational Linguistics.

Xu, Dongfang et al. 2020. »Multi-class Hierarchical Question Classification for Multiple Choice Science Exams«, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, hrsg. v. Calzolari, Nicoletta et al., S. 5370–5382. Marseille, France: European Language Resources Association.

Xu, Jun; Li, Hang 2007. »AdaRank: a boosting algorithm for information retrieval«, in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, S. 391–398. New York, NY, USA: Association for Computing Machinery.

Yang, Eugene; Lewis, D.; et al. 2018. »Retrieval and Richness when Querying by Document«.

Yang, Yinfei et al. 2021. »Neural Retrieval for Question Answering with Cross-Attention Supervised Data Augmentation«, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, S. 263–268. Online: Association for Computational Linguistics.

Yedidia, Adam B. 2016. »Against the F-score«.

Zhan, Jingtao et al. 2020. »An Analysis of BERT in Document Ranking«, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, S. 1941–1944. Virtual Event China: ACM.

Zhang, Dell; Lee, Wee Sun 2003. »Question classification using support vector machines«, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, S. 26–32. New York, NY, USA: Association for Computing Machinery.

Zhang, Yue et al. 2023. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. arXiv.

Zhao, Ruochen et al. 2023. »Retrieving Multimodal Information for Augmented Generation: A Survey«, in *Findings of the Association for Computational Linguistics: EMNLP 2023*, hrsg. v. Bouamor, Houda; Pino, Juan; Bali, Kalika, S. 4736–4756. Singapore: Association for Computational Linguistics.

Zhou, Qingyu et al. 2018. »Neural Question Generation from Text: A Preliminary Study«, in *Natural Language Processing and Chinese Computing*, hrsg. v. Huang, Xuanjing et al., S. 662–671. Cham: Springer International Publishing.

Zhu, Fengbin et al. 2021. *Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering*. arXiv.

Zhu, Wenhao et al. 2023. »A Hybrid Text Generation-Based Query Expansion Method for Open-Domain Question Answering«, in *Future Internet* 15, 5, S. 180.

Erklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit mit dem Thema: *Explorative Untersuchung der Effektivität semantischer Klassifizierung in der Abrufung für RAG-Systeme* selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

_____ _____

(Ort, Datum) (Unterschrift)

Anlage „Erklärung zur Verwendung generativer KI-Systeme“

Bei der Erstellung der eingereichten Arbeit habe ich die nachfolgend aufgeführten auf künstlicher Intelligenz (KI) basierten Systeme benutzt:

1. GPT-4 über https://chat.openai.com
2. Elicit über elicit.com

Ich erkläre, dass ich

- mich aktiv über die Leistungsfähigkeit und Beschränkungen der oben genannten KI-Systeme informiert habe,[304]
- die aus den oben angegebenen KI-Systemen direkt oder sinngemäß übernommenen Passagen gekennzeichnet habe,[305]
- überprüft habe, dass die mithilfe der oben genannten KI-Systeme generierten und von mir übernommenen Inhalte faktisch richtig sind,
- mir bewusst bin, dass ich als Autorin bzw. Autor dieser Arbeit die Verantwortung für die in ihr gemachten Angaben und Aussagen trage.

Die oben genannten KI-Systeme habe ich wie im Folgenden dargestellt eingesetzt:

Arbeitsschritt in der wissenschaftlichen Arbeit[306]	**Eingesetzte(s) KI-System(e)**	**Beschreibung der Verwendungsweise**
Definierung der Zielsetzung / Literaturrecherche	Elicit	Elicit wurde verwendet, um explorativ auf interessante Fragen relevante Quellen und Antworten abzufragen. Zudem wurde Elicit verwendet, um vor der offiziellen Literaturrecherche zu verschiedenen Fragen aufzuzeigen ob überhaupt zu der spezifischen Fragestellung Literatur vorhanden ist oder Begriffe eventuell umformuliert werden müssen.
Verfassung der Arbeit	GPT-4	Es wurde gelegentlich Chat-GPT mit GPT-4 als Wörterbuch verwendet um die Korrekte Rechtschreibung von Wörtern, Synonyme und die korrekte Grammatikalische Form für einzelne Ausdrücke erfragt.
Verfassung der Arbeit	GPT-4	Das Abkürzungsverzeichnis wurde nach Vervollständigung mithilfe von Chat-GPT in Alphabetischer Reihenfolge sortiert.

Ort, Datum, Unterschrift

[304] U.a. gilt es hierbei zu beachten, dass an KI weitergegebene Inhalte ggf. als Trainingsdaten genutzt und wiederverwendet werden. Dies ist insb. für betriebliche Aspekte als kritisch einzustufen.
[305] In der Fußnote Ihrer Arbeit geben Sie die KI als Quelle an, z.B.: Erzeugt durch Microsoft Copilot am dd.mm.yyyy. Oder: Entnommen aus einem Dialog mit Perplexity vom dd.mm.yyyy. Oder: Absatz 2.3 wurde durch ChatGPT sprachlich geglättet.
[306] Beispiele hierfür sind u.a. die folgenden Arbeitsschritte: Generierung von Ideen, Konzeption der Arbeit, Literatursuche, Literaturanalyse, Literaturverwaltung, Auswahl von Methoden, Datensammlung, Datenanalyse, Generierung von Programmcodes