Ansätze für einen frageorientierten Zugang zu Studienordnungen

Masterarbeit

im Studiengang Computing in the Humanities der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Lehrstuhl für Medieninformatik

Verfasser: Frauke Möglich

Prüfer: Prof. Dr. Andreas Henrich

Bamberg 2026

URN: urn:nbn:de:bvb:473-irb-112943x

DOI: https://doi.org/10.20378/irb-112943

Zusammenfassung

Ziel dieser Arbeit war es, einen frageorientierten Zugang zu Studienordnungen für Studierende zu entwickeln. Dabei wurden verschiedene Ansätze untersucht, die es Studierenden ermöglichen, anhand typischer Fragestellungen relevante Stellen in ihren Studienordnungen zu finden und so den Zugang zu Informationen zu erleichtern. Zu Beginn wurde das Informationsbedürfnis der Zielgruppe analysiert und eine nutzenden-orientierte Evaluationsmetrik entwickelt. Anschließend wurden unterschiedliche Methoden – darunter TF-IDF, BERT-basierte Modelle sowie das Large Language Model Mistral Large – getestet und bewertet. Die Ergebnisse zeigten, dass unter bestimmten Bedingungen TF-IDF die besten Resultate lieferte, gefolgt von Mistral und den BERT-basierten Modellen. Es wurden weiterhin verschiedene Merkmale der Texte untersucht, mit dem Ziel, die Bedingungen, in denen die Ansätze gut oder schlecht funktionieren, herauszufinden. Die Ergebnisse, die mit Mistral Large gewonnen wurden, wurden außerdem nach Fehlern, die typisch für Mistral Large sein könnten, hin untersucht. Explorativ wurde ein BERT-basiertes Modell unter Verwendung von durch TF-IDF gefundenen Features getestet.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Motivation & Leitfragen

1.1 Motivation der Arbeit

Ein Studium verspricht häufig die beste Zeit des Lebens (Appenzeller and Kersting [2013]) zu sein, bringt jedoch auch viele Herausforderungen mit sich: So zeigt sich, dass das Stresserleben Studierender in Deutschland auf einem hohen Niveau ist, nach Herbst et al. [2016] gaben 53% der Teilnehmenden ein hohes Stresserleben an. Ursachen der Belastung sind unterschiedlicher Natur, neben Stressoren, die sich aus den inhaltlichen Anforderungen eines Studiums oder der finanziellen Situation ergeben (Sendatzki and Rathmann [2022]) wird auch die Angst vor der Überschreitung der Regelstudienzeit angegeben (Herbst et al. [2016]). Neben anderen studiumsrelevanten Informationen finden Studierende die Bestimmungen zur Regelstudienzeit in ihren Studien- und Fachprüfungsordnungen (im Folgenden oft Ordnungen oder Studienordnung genannt). Hier werden sie vor weitere Herausforderungen gestellt: Das Ziel von rechtlichen Texten ist eher selten ihre zielgruppenverständliche Formulierung (Helmchen [2017]). Zusätzlich befinden sich in den Ordnungen auch viele Informationen, die nicht direkt relevant für Studierende sind, sondern der rechtlichen Absicherung der Universität dienen.

Methoden im Bereich des Natural Language Processing (NLP) bieten Ansätze, mit Texten auf verschiedene Art und Weise zu interagieren (Min et al. [2023]) und somit Chancen, Studierende bei der Interaktion mit ihren Studienordnungen zu unterstützen. Es stellt sich daher die Frage, ob und inwiefern Methoden aus dem Bereich NLP genutzt werden können, um Studierenden den Zugang zu ihren Studienordnungen zu erleichtern und ihr Informationsbedürfnis zu erfüllen.

Die Formulierung des eigenen Informationsbedürfnisses fällt je nach Person unterschiedlich aus (Aula [2003]) und der oder die Suchende ist sich nicht darüber klar, wie genau das System funktioniert und welche Ausdrücke zu welchen Ergebnissen und warum führen (Belkin [2000]). Daher liegt es nahe, dass Informationssysteme Handlungsempfehlungen geben sollten, um Suchenden zu helfen, ihre Probleme besser zu verstehen und die Ressourcen des Systems effektiver zu nutzen (Belkin [2000]). In dieser Arbeit wurde ein frageorientierter Zugang gewählt, um Studierenden Handlungsempfehlungen zu präsentieren.

Das Ziel dieser Arbeit ist daher, mögliche Lösungsansätze herauszuarbeiten, um das Informationsbedürfnis Studierender bezüglich ihrer Studienordnungen mit Hilfe eines frageorientierten Zugangs optimal zu erfüllen und dabei die Möglichkeiten der verschiedenen Ansätze zu evaluieren.

Aus diesen Überlegungen ergeben sich zwei thematische Bereiche mit eigenen Leitfragen:

1. Zielgruppenpassung

- Wie lässt sich ein Informationssystem im Zusammenhang mit Studienordnungen aufbauen, das den Bedürfnissen der Zielgruppe entspricht?
  Daraus folgen weitere Leitfragen:
  1. Welchen Inhalt hat das Informationsbedürfnis der Zielgruppe im Kontext von Studien- und Fachprüfungsordnungen?
  2. Mit welcher Performanz eines möglichen Informationssystems wäre die Zielgruppe zufrieden?
  3. Wie kann ein Informationssystem für die Zielgruppe aufgebaut sein?

2. Methoden

- Welche Ansätze eignen sich inwiefern, um ein solches Informationssystem aufzubauen?
  Daraus folgen weitere Leitfragen:
  1. Welche Ansätze kommen für den gegebenen Kontext in Frage?
  2. Wie kann die Performanz der verwendeten Ansätze evaluiert werden?
     (a) Hinsichtlich der Vergleichbarkeit mit anderen Information Retrieval Systemen?
     (b) Hinsichtlich der Passung zu den Bedürfnissen der Zielgruppe?
  3. Können bestimmte Eigenschaften der Anfragen genutzt werden, um die Performanz der Ansätze vorherzusagen und dadurch eine effektive Kombination zu ermöglichen?

1.2 Aufbau der Arbeit

Im Grundlagenteil (Kapitel 2) wird der Forschungskontext sowie die vorhandenen Methoden, Metriken und Theorien erläutert und im Konzeptteil (Kapitel 3) wird dargelegt, wie das System aufgebaut ist und warum bestimmte Entscheidungen getroffen wurden. Anschließend wird im Kapitel 4 Umsetzung die Implementierung der verwendeten Ansätze kurz dargestellt und im Ergebnisteil (Kapitel 5) werden die Ergebnisse aus den Vorbereitungen, aus der Anwendung der Ansätze sowie aus der Nachbereitung berichtet. In Kapitel 6 werden die Ergebnisse diskutiert, es wird auf offene Fragen eingegangen und die Beantwortung der Leitfragen wird abschließend betrachtet.

Kapitel 2

Grundlagen & Kontext

2.1 Information Retrieval (IR)

Information Retrieval (IR) ist definiert als die Wissenschaft der Suche nach Informationen in Dokumenten, nach Dokumenten selbst und nach Metadaten, die Daten beschreiben, sowie nach Datenbanken von Tönen und Bildern und ein Anwendungsgebiet des maschinellen Lernens (Shinde and Shah [2018]). Es befasst sich traditionell mit der Darstellung, Speicherung, Suche und Auffindung von Informationen, die für Nutzende relevant sind (Ingwersen [1992]) und gehört zu einer der ältesten Disziplinen der Informatik (Arora et al. [2016]). IR ist somit ein Teilgebiet des Natural Language Processing (NLP) (Strzalkowski [1995]). NLP bezeichnet die Verwendung computergestützter Methoden zur Verarbeitung gesprochener oder geschriebener unstrukturierter Texte, die als gängiges Kommunikationsmittel des Menschen dienen (Assal et al. [2011]).

Heutzutage ist IR aufgrund der Integration in viele alltägliche Anwendungen, wie z.B. die Internetsuche oder die Verwendung von Chatbots, ein wichtiger Bestandteil unseres täglichen Lebens geworden (Hambarde and Proença [2023]). Dabei spielt die Interaktion zwischen den Nutzenden und den Informationen eine wichtige Rolle für ein effektives Systemdesign (Radlinski and Craswell [2013]). Das Ziel ist, Ergebnisse zu finden und darzubieten, die möglichst stark mit der Anfrage des Nutzenden verbunden sind (Hambarde and Proença [2023]). Häufig ist die Anzahl an möglichen Treffern bzw. Dokumenten sehr groß, daher arbeiten viele Systeme zweistufig: Zunächst werden mithilfe eines Retrievers relevante Dokumente gefunden, um diese in einem zweiten Schritt durch einen Reranker nach Relevanz für die Anfrage zu sortiert (Hambarde and Proença [2023]).

Die Methoden, die dazu verwendet werden können, werden grundsätzlich in überwachte und nicht überwachte Methoden unterschieden (Mitra and Craswell [2018]). Für überwachte Methoden werden in der Regel Trainingsdaten benötigt, die häufig nicht zur Verfügung stehen oder nur sehr aufwendig erarbeitet werden können (Ding et al. [2020]). Der Fokus dieser Arbeit liegt aus diesen Gründen auf Methoden, die in ihrer Anwendung keine Trainingsdaten benötigen.

2.1.1 Methoden des Information Retrieval (IR)

Die Grundlage für die Entwicklungen im Information Retrieval legte der 1945 veröffentlichte Artikel *As We May Think* von Vannevar Bush. Darin beschreibt er mit

„Memex“ ein Gerät, das einen automatisierten Zugang zu einer großen Menge gespeicherter Informationen ermöglichen sollte (Bush et al. [1945]).

Die ersten Information Retrieval Systems basierten auf Booleans und ermöglichten Nutzenden, ihre Suche durch eine komplexe Kombination von ANDs, ORs und NOTs zu realisieren (Singhal et al. [2001]). Im Vector Space Model (VSM) wurde erstmals ein Ranking über die Berechnung der Ähnlichkeit zwischen der Anfrage und den möglichen Treffern möglich; dies geschieht durch Repräsentation der Texte als Vektoren im mehrdimensionalen Raum (Salton [1975]). In frühen Anwendungen des VSM wird jedem Term (das könnten z.B. Wörter sein) des Vokabulars eine Dimension in einem hochdimensionalen Vektorraum zugewiesen, und falls ein Term im betrachteten Text oder der Anfrage vorkommt, wird der Dimension ein i.d.R. nicht-negativer Wert zugewiesen (Singhal et al. [2001]). Diese Art der traditionellen Anwendung im VSM verwendet somit dünn besetzte Vektoren, die bei einem großen Vokabular in vielen ihrer Dimensionen mit Nullen gefüllt sind (Luan et al. [2021]). Zu diesen traditionellen Ansätzen zählen Modelle, die auf Bag-of-Words-Repräsentationen und einer gewichteten Termfrequenz-Inversen-Dokumentfrequenz (TF-IDF) basieren (Mandikal and Mooney [2024]) und bieten eine hohe Effizienz und eine robuste Leistung (Li et al. [2025a]).

Dagegen wird in neuere Ansätze unterschieden, die mit dichten Vektoren arbeiten (vergleiche Luan et al. [2021]). Dieses neue Paradigma entwickelte sich weiter mit dem Aufkommen von Deep Learning Methoden im Information Retrieval, was die Verwendung von dichten Vektoren ermöglichte (Mandikal and Mooney [2024]). Deep Learning ist eine Unterkategorie des maschinellen Lernens, die sich auf künstliche neuronale Netze mit vielen Schichten, sogenannte tiefe neuronale Netze, konzentriert (LeCun et al. [2015]). Im Bereich des Deep Learning hat sich das Vortraining mit selbstüberwachtem Lernen auf der Grundlage umfangreicher, nicht-gelabelter Daten zum vorherrschenden Ansatz (Transferlernen) entwickelt (Wang et al. [2023a]). Die Idee des Transferlernens besteht darin, das aus anderen Aufgaben gewonnene Wissen wiederzuverwenden und auf neue Aufgaben anzuwenden (Wang et al. [2023a]). Ein Vorteil der durch die neueren Deep Learning Methoden im Vergleich zu den traditionellen Ansätzen liegt in der Vereinigung der Schritte der Merkmalsextraktion und Klassifikation, die in den traditionellen Ansätzen getrennt waren (Perumal et al. [2024]). Prominente Vertreter dieser Gruppe sind beispielsweise Word2Vec (Mikolov et al. [2013]), GloVe (Pennington et al. [2014]) und Autoencoder (Hinton and Salakhutdinov [2006]). Allerdings gilt hierbei nicht immer automatisch, dass mit dichten Vektoren bessere Ergebnisse erzielt werden als mit dünn besetzten Vektoren (Luan et al. [2021]).

Mit der Einführung der Transformer-Architektur 2017 im Artikel *Attention is All You Need* (Vaswani et al. [2017]) entstanden neue Möglichkeiten: Im Gegensatz zu früheren Architekturen wie RNNs (Recurrent Neural Networks) und CNNs (Convolutional Neural Networks) setzt die Transformer-Architektur ausschließlich auf den Attention-Mechanismus, dies ermöglicht eine starke Parallelisierung während des Trainings, wodurch sehr große Trainingsdaten verwendet werden können (Vaswani et al. [2017]). Die Modelle werden mit umfassenden und allgemeinen Textkorpora vortrainiert und daher PLMs (Pre-trained Language Models) genannt (Jiang and Cai [2024]). Durch ihr Vortraining können sie eine große Menge semantischen Wissens erfassen und somit Textinhalte vergleichsweise gut verstehen (Zhao et al. [2024]). Während Vaswani et al. [2017] eine Encoder-basierte und Decoder-basierte Architektur beschreiben, basieren spezifische Modelle in der Regel ent-

weder auf der einen oder der anderen Architektur je nach Anwendungsgebiet (Liu et al. [2023]). Wichtige Vertreter der PLMs sind neben anderen BERT-basierten Modellen (Bidirectional Encoder Representations from Transformers, Devlin et al. [2019]), sind GPT (Generative Pre-trained Transformer, Radford et al. [2018]), XLNet (Yang et al. [2019]) und T5 (Text-to-Text Transfer Transformer, Raffel et al. [2020]).

Mit wachsender Anzahl an Parametern gehen (P)LMs in Large Language Models (LLMs) über, der Übergang ist quantitativ begründet, aber qualitativer Natur bezüglich der Fähigkeiten eines LLMs (Zhu et al. [2024]). Denn mit wachsender Größe der Modelle wächst auch ihre Performanz (Kaplan et al. [2020]) und das Paradigma des „pre-train and fine-tune “ wird abgelöst durch „pre-train, prompt, and predict “ (Liu et al. [2023]). Außerdem wird das Vorgehen des Rankings der Dokumente hin zur Generierung einer Antwort auf die Anfrage verändert (Djeddal et al. [2024]). Hierbei spielt Prompting eine entscheidende Rolle, da es im Zero-Shot Learning das Verhalten der LLMs steuert (Liu et al. [2023]). Die gesteigerte Performanz im Zero- oder Few-Shot Learning lässt sich durch das Scaling Law begründen, nach dem insbesondere die Modellgröße, die Größe des verwendeten Trainingsdatensatzes und die Rechenleistung, gemessen in Floating Point Operations per Second (FLOPs), die Performance des Modells beeinflussen, während Architekturunterschiede wie beispielsweise die Netzwerkbreite und -tiefe eine geringere Rolle spielen (Kaplan et al. [2020]). Als erster bedeutender Vertreter der LLMs kann GPT-3 (Brown et al. [2020]) gelten, das mit 175 Milliarden Parametern etwa zehnmal größer ist als seine Vorgänger unter den dichten Sprachmodellen und dessen Architektur sich im Vergleich zum vorherigen Modell GPT-2 kaum geändert hat (Brown et al. [2020]). Zu den weiteren Vertretern der LLMs gehören unter anderem Llama (Touvron et al. [2023]) von Meta AI, die Modelle von Mistral oder die multi-modalen Modelle Gemini (Team et al. [2023]) von Google DeepMind.

2.1.2 Question Answering (QA) als Teilbereich des IR

Question Answering (QA) ist ein spezialisierter Bereich des IR, in dem Forschung aus verwandten Bereichen wie des IR, der Information Extraction (IE) und des NLP kombiniert wird (Allam and Haggag [2012]). IE-Systeme nehmen Texte in natürlicher Sprache, d.h. in unstrukturierter Form, als Eingabe und erzeugen strukturierte Informationen, die nach bestimmten Kriterien spezifiziert sind und für eine bestimmte Anwendung relevant sind (Singh [2018]). Im Gegensatz zum reinen IR, bei dem lediglich die Dokumente nach Relevanz geordnet werden und der/die Nutzende mit dem Dokument oder mit einer Passage des Textes selbst die Anfrage beantworten muss, ist das Ziel des QA, genaue Antworten anstelle von Dokumenten bzw. Textpassagen zu liefern (Allam and Haggag [2012]). Somit kann QA als eine fortgeschrittene Form des IR aufgefasst werden (Cao et al. [2010]). Da Nutzende sich häufig eine präzise Antwort auf ihre Anfrage wünschen (Hirschman and Gaizauskas [2001]), wächst die Nachfrage nach QA-Systemen (Pudaruth et al. [2016]).

Die Anfragen, die Nutzende stellen, lassen sich nach Kolomiyets and Moens [2011] in *Faktfrage*, *Listenfragen*, *Definitionsfragen* und *komplexe Fragen* einteilen. Besonders häufig werden Faktfragen in QA-Systemen gestellt (Allam and Haggag [2012]). Neben der Klassifizierung der Anfragen der Nutzenden ist es ebenfalls wichtig, die verschiedenen Arten von QA-Systemen zu verstehen: Grundsätzlich lassen sich QA-Systeme in domäneoffene und domäne-spezifische Systeme einteilen, bei den ersteren sind Anfragen bezüglich nahezu jedem Thema möglich, während letztere nur Anfragen zu spezifischen Themen

zulassen (Allam and Haggag [2012]).

Ein typisches QA-System besteht aus drei verschiedenen Modulen, 1. das *Query processing Module*, um die Anfrage zu klassifizieren, 2. das *Document Processing Module*, für das Information Retrieval sowie 3. das *Answer Processing Module*, um eine Antwort zu extrahieren (Allam and Haggag [2012]). Im *Query processing Module* kommt es neben einer Klassifizierung der Anfrage hinsichtlich verschiedener Aspekte auch zu einer Reformulierung der Anfrage, um die Performanz im *Document Processing Module*, also dem IR, zu verbessern (Allam and Haggag [2012]).

2.1.3 Evaluierung im Information Retrieval

Unter Evaluierung versteht man die Bewertung der Leistung oder des Wertes eines Systems, Prozesses, Produktes oder Richtwertes und somit ist Evaluierung eine unverzichtbare Notwendigkeit in der Wissenschaft, der Technik und vielen anderen Bereichen (Saracevic [1995]). Dabei ist es immer ein heikles Thema, was genau die Grundlage für eine Evaluation bilden sollte (Saracevic [1995]). In der Evaluation von IR-Systemen bildet häufig die Relevanz der Ergebnisse des IR-Systems diese Grundlage und ist ein viel diskutierter Aspekt des IR (Robertson and Hancock-Beaulieu [1992]): Um zu wissen, welche Treffer für eine bestimmte Anfrage als relevant gelten, benötigt es ein Verständnis davon, wie Relevanz definiert werden kann. Obwohl das Konzept der Relevanz im IR fundamental ist, fehlt ein gemeinsames Verständnis über die Definition und die Verwendung in der Evaluation von IR-Systemen (Kagolovsky and Möhr [2001]).

Um die Effektivität von IR-Systemen zu bewerten, können zwei verschiedene Ansätze verfolgt werden: nutzenden-orientierte Ansätze und system-orientierte Ansätze (Moghadasi et al. [2013]). Betrachtet man system-orientierte Ansätze, so wird Relevanz als eine Eigenschaft des Systems betrachtet (Cosijn and Ingwersen [2000]). Der nutzenden-orientierte Ansatz fokussiert sich auf die Zufriedenheit der Nutzenden durch die Beobachtung ihrer Interaktionen mit dem System und ist somit aufwändiger als der system-orientierte Ansatz (Fidel [1993]). Da viele der modernen IR-Systeme immer mehr im Alltag integriert sind und persönlicher werden, gibt es kein objektives Kriterium in der Bewertung der Relevanz der Ergebnisse mehr. Immer mehr muss die aktuelle Situation der suchenden Person sowie vorherige Interaktionen mit dem IR-System berücksichtigt werden (Hofmann et al. [2016]) und der nutzende-zentrierte Ansatz rückt somit in den Vordergrund.

Werden Treffer nach Relevanz bewertet, so kann dies in binärer Form (relevant vs. non-relevant) oder abgestufter Form (z.B. non-relevant, marginally relevant, fairly relevant, highly relevant) geschehen (Kekäläinen [2005]). Die Wahl zwischen binärer oder abgestufter Relevanz hat zur Konsequenz, welche Evaluationsmethoden angewendet werden können (Kekäläinen [2005]).

Die Evaluationsmethoden haben sich im Laufe der Zeit weiterentwickelt und umfassen sowohl offline als auch online Ansätze (Voorhees [2019]): Traditionell werden Suchsysteme offline durch die Erstellung von Testsammlungen, die aus einer Sammlung von Dokumenten, einer Reihe von Themen und einer Reihe von Relevanzbewertungen, die angeben, für welches Thema ein Dokument relevant ist, bewertet, bestehen (Fu et al. [2022]). Dieses Vorgehen wird häufig als Cranfield-Paradigma bezeichnet (Fu et al. [2022]). Die Cranfield-Paradigmen haben sich als eine effektive Bewertung insbesondere für die Qualität von Ranglisten-Suchfunktionen erwiesen, jedoch skalieren sie nicht gut für große Websuchen (Radlinski et al. [2008]). Bereiche, die sich gut für offline Evaluation eignen, sind solche, in

denen die Relevanz der Dokumente zuverlässig und unverzerrt durch Expert*Innen oder repräsentative Nutzende erfolgen kann (Radlinski et al. [2008]). Ein anderer Ansatz ist die Online-Evaluation, bei der das IR-System Nutzenden präsentiert wird und ihre Interaktionen mit dem System beobachtet werden (Hofmann et al. [2016]). Dadurch, dass echte Personen das IR-System benutzen, kann Online-Evaluation als eine Ergänzung zur Offline-Evaluation genutzt werden, deren Ergebnisse häufig leichter zu interpretieren, aber weniger realistisch sind (Hofmann et al. [2016]). Einen Überblick über die Vielzahl verschiedener Metriken bieten zum Beispiel Bama et al. [2018] oder Zuva and Zuva [2012].

Kapitel 3

Konzeptteil & Methoden

3.1 Aufbau des Systems

Häufig wird in IR-Systemen mit einem zweistufigen System gearbeitet: Einem Retriever und einem Reranker (Hambarde and Proença [2023]). Dabei ist das Ziel des Retrievers, mit ressourcenarmen Methoden eine erste Auswahl an Dokumenten zu finden (Guo et al. [2022]). Der Reranker kann dann in weiteren Schritten die in der vorherigen Stufe erstellte Rangliste der Dokumente bereinigen und verbessern (Guo et al. [2022]). Die Verwendung eines Retrievers und eines Rerankers ergibt somit vor allem in Anwendungen Sinn, bei denen es eine große Anzahl möglicher Treffer gibt. Da dies im vorliegenden Kontext nicht der Fall ist, wurde auf eine entsprechende Einteilung verzichtet.

Je nach verwendeter Methode ergibt sich ein etwas anderer Aufbau des Systems. Allen gemein ist die Aufbereitung der Studienordnungen sowie das Erstellen der Anfragen: Sowohl die Anfragen als auch die Textstellen sollen mit Hilfe der verschiedenen Ansätze verarbeitet werden und anschließend sollen die Ergebnisse ausgewertet werden. Die Studien- und Fachprüfungsordnungen sollen in a) Paragrafen und b) Subparagrafen eingeteilt und c) die Titel der Paragrafen ausgelesen werden, dazu sollen Reguläre Ausdrücke verwendet werden. Um die Anfragen zu gewinnen, soll zunächst das Informationsbedürfnis der Zielgruppe untersucht werden, und als Typical Asked Questions (TAQs) zusammengefasst werden. In einem weiteren Schritt werden die TAQs dann, je nach Art der Anwendung, in die Anfragen umgewandelt. Für das Ranking und die Evaluation von TF-IDF und BERT-basierten Methoden sollen in einem ersten Schritt die Kosinus-Ähnlichkeiten der Textstellen-Anfragen-Paare berechnet und dann entsprechend absteigend sortiert werden. In einem weiteren Schritt soll geprüft werden, an welcher Stelle die tatsächlich relevante Textstelle steht, und die Evaluationsmetriken sollen berechnet werden. Das Anwenden von dem LLM mistral-large-2411 geschieht nach Erstellen eines Prompts über eine API-Anfrage. Hier soll manuell ausgewertet werden und nach Fehlerursachen gesucht werden. Das Vorgehen ist in Abbildung 1 dargestellt.

3.2 Typical Asked Questions (TAQs) und Umfrage 1

IR-Systeme müssen sich auch dem Problem stellen, wie die Suche spezifiziert und die Interaktion zwischen den Produzierenden von Informationen und den Nutzenden von Informationen gestaltet werden kann (Saracevic [1995]). Im Fokus steht dabei das Informati-

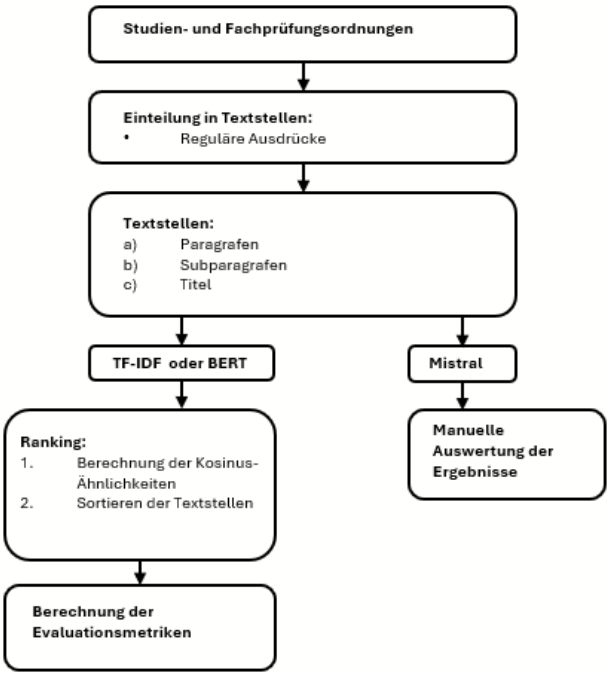

Abbildung 1: *Pipeline zur Extraktion und Bewertung relevanter Textstellen mithilfe von TF-IDF, BERT und Mistral.*

onsbedürfnis, das es zu identifizieren gilt, auch um ein IR-System anschließend evaluieren zu können (Arora et al. [2016]).

Bemerkenswerterweise bekommen Nutzende mit demselben Informationsbedürfnis, aber unterschiedlich formulierten Anfragen, häufig unterschiedliche Ergebnisse sowohl bezüglich der Art der Ergebnisse als auch bezüglich ihrer Güte (Penha et al. [2022]). Im vorliegenden Kontext soll dies verhindert werden, indem für jedes gefundene Informationsbedürfnis genau eine vorformulierte Frage bereitgestellt werden soll. Das bedeutet, dass eine Zielgruppe mit ähnlichen Informationsbedürfnissen aus einem Pool möglicher Anfragen je nach ihrem Informationsbedürfnis auswählen könnte, um sicherzustellen, dass das Informationsbedürfnis erfüllt wird und nicht an der Art und Weise der Formulierung der Anfrage scheitert. Daher wurde sich dazu entschieden, die Informationsbedürfnisse Studierender durch eine Umfrage herauszufinden, mit dem Ziel, sie als Typical Asked Questions (TAQs) zu strukturieren.

Um das System anhand von TAQs aufzubauen, werden bestimmte Annahmen getroffen. So wird angenommen, dass Studierende häufig vor ähnliche Herausforderungen während des Studiums gestellt werden und Studien- und Prüfungsordnung nur eine begrenzte Menge an für Studierende relevanten Informationen enthalten. Daher wird in dieser Arbeit davon ausgegangen, dass es eine abgrenzbare Menge an typischen Fragen seitens Studierender gibt und dass diese mit Hilfe einer Umfrage gefunden werden können. Um möglichst viele TAQs zu identifizieren, sollen Items verwendet werden, die verschiedene Aspekte fokussieren. Leider kann dabei trotzdem nicht sichergestellt werden, dass wirklich alle Fragen seitens Studierender herausgefunden werden; es wird sich in diesem Sinne auf die typischen Fragen beschränkt.

So können Studierende im besten Fall die für sie passende TAQ auswählen, um ihr Informationsbedürfnis zu befriedigen. Normalerweise sollten IR-Systeme jedoch einen hohen Grad an Interaktivität begünstigen (Radlinski and Craswell [2013]), da sich das Informationsbedürfnis im Suchprozess durch die gefundenen Ergebnisse ändern kann (Belkin et al. [1993]). Im vorliegenden Kontext soll die Interaktivität durch die Möglichkeit der Suchenden, die verschiedenen TAQs auswählen zu können, angeboten werden. Außerdem ist davon auszugehen, dass, falls die passende Textstelle gefunden wurde, das Informationsbedürfnis der Suchenden befriedigt wurde und keine weiteren Interaktionen nötig sind.

3.3 Methodenauswahl

Die Entwicklung des IR hat verschiedene Wellen durchlaufen, von ihren Anfängen mit term-basierten Methoden (Guo et al. [2018]), über die Einführung von Vektorraummodellen (Raghavan and Wong [1986]) hin zu transformer-basierten Modellen wie Pretrained Language Models (PLM) und Large Language Models (LLMs) (Zhu et al. [2023]). In der vorliegenden Arbeit soll je ein Ansatz aus den traditionellen Ansätzen, den transformerbasierten PLMs und LLMs angewendet werden. Dabei werden nur Ansätze in Betracht gezogen, die im vorliegenden Kontext, d.h. insbesondere ohne Trainingsdaten, anwendbar sind und kostenlos nutzbar sind. Von den traditionellen Ansätzen kommen vor allem TF-IDF und BM25 für den gegebenen Anwendungskontext in Frage.

3.3.1 Traditionelle Ansätze

Abwägung von TF-IDF und BM25 Sowohl TF-IDF (Sparck Jones [1972]) als auch BM25 (Robertson et al. [1995]) bauen auf Bag-of-Words auf: Die zentrale Idee von TF-IDF ist, dass Wörter, die in einem Dokument häufiger auftauchen, wichtig sind für den Kontext des Dokuments (Marwah and Beel [2020]). Dabei sollen jedoch Wörtern, die über alle Dokumente betrachtet seltener vorkommen, eine höhere Gewichtung gegeben werden, als solchen, die insgesamt häufig vorkommen, wie beispielsweise Artikel oder Pronomen (Marwah and Beel [2020]). Durch diese Eigenschaft, die relative Wichtigkeit von Begriffen in einem Dokument oder Korpus aufzeigen zu können, ist TF-IDF weit verbreitet trotz Nachteilen wie den Curse of Dimensionality[1], die Data Sparsity[2] und die Schwierigkeit, Bedeutung zwischen Wörtern in einem Dokument zu erfassen (Abubakar et al. [2022]).

Diese Problem tritt insbesondere bei Datensätze, die auf Grund des häufigen Gebrauches seltener Wörter besonders dünn besetzte Vektoren produzieren, auf (vergleiche Saif et al. [2014]). Der vorliegende Datensatz ist relativ klein und es muss geprüft werden, wie häufig seltene Wörter verwendet werden. In diesem Zusammenhang kann es bei der Verwendung von traditionellen, term-basierten Methoden auch zu einem Vocabulary Mismatch kommen, wenn in der Anfrage und den Dokumenten unterschiedliche Wörter verwendet werden (Guo et al. [2022]).

[1] Der Curse of Dimensionality beschreibt das Problem, dass bei einer hohen Anzahl an Dimensionen bei der Anwendung von Ähnlichkeitsberechnungen wie beispielsweise der Kosinus-Ähnlichkeit einen Verlust an aussagekräftiger Unterscheidung zwischen ähnlichen und unähnlichen Objekten zu beobachten ist (Assent [2012]).

[2] Unter Data Sparsity versteht man das Problem, dass die durch TF-IDF erzeugten Vektoren dadurch, dass in einem Dokument nur ein kleiner Teil aller im Vokabular enthaltenen Begriffe vorkommt, relativ dünn besetzt sind (d. h. viele Nullen enthalten) bei gleichzeitig hoher Dimensionalität.

BM25 baut auf TF-IDF auf und ist ein Algorithmus zum Sortieren der Dokumente, der ebenfalls wie TF-IDF statistische Eigenschaften wie die Termhäufigkeit, Dokumenthäufigkeit und Dokumentlänge verwendet (Rosa et al. [2021]). Im Gegensatz zu TF-IDF führt BM25 jedoch eine Längennormalisierung durch, um Unterschiede in der Dokumentlänge zu berücksichtigen (Rosa et al. [2021]). Außerdem gibt es anders als bei TF-IDF zwei frei wählbare Parameter, die zur Verbesserung der Ergebnisse angepasst werden können (Rosa et al. [2021]). BM25 bietet somit einige Vorteile gegenüber TF-IDF. Im Anwendungskontext von Studienordnungen ist jedoch fraglich, ob diese Vorteile greifen: So sind die Paragrafen zwar unterschiedlich lang, jedoch kann vermutet werden, dass der Informationsgehalt nicht abnimmt, da der Schreibstil vermutlich konstant ist und daher wäre eine Normalisierung auf Grund der Dokumentenlänge nicht zielführend. Beispielsweise Kadhim [2019] haben sowohl TF-IDF als auch BM25 genutzt, um Schlüsselwörter aus Daten zu extrahieren, die auf Twitter gesammelt wurden: Hier hat die TF-IDF als Methode einen besseren F1-Score erzielen können. Diese Punkte sprechen insgesamt dafür, TF-IDF als Vertreter der term-basierten Methoden zu verwenden.

Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency Inverse Document Frequency (TF-IDF) ist ein relativ altes, aber effektives und einfaches Verfahren, um die Relevanz bestimmter Terme für ein Dokument (oder Textabschnitt) zu bestimmen (Sparck Jones [1972]). Dabei wird sowohl beachtet, wie häufig ein Term im betrachteten Dokument vorkommt, als auch wie häufig der Term über alle Dokumente betrachtet vorkommt. TF-IDF wird als statistisches Verfahren kategorisiert, obwohl es deterministisch ist (Ramos et al. [2003]). Das Resultat ist die Multiplikation der Term Frequency und der Inverse Document Frequency:

Term Frequency (TF) misst, wie oft ein bestimmter Term in einem Dokument, im Fall der Studienordnungen könnte es z.B. ein Paragraf sein, vorkommt. Zum Beispiel könnte der Term „Anmeldung“ im betrachteten Paragrafen zweimal vorkommen. Da Paragrafen (bzw. Textstellen im Allgemeinen) nicht zwingend gleich lang sind, bietet es sich an, die relative Häufigkeit zu berechnen.

$$\mathrm{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Dabei gilt:

- t ist ein Term (Wort oder Token),
- d ist ein Dokument aus einer Sammlung von Dokumenten,
- $f_{t,d}$ ist die Häufigkeit des Terms t im Dokument d,
- $\sum_{t' \in d} f_{t',d}$ ist die Gesamtanzahl aller Terme im Dokument d.

Die Inverse Term Frequency gibt den Termen, die häufiger vorkommen, eine geringere Gewichtung im Vergleich zu selten auftretenden Termen (Qaiser and Ali [2018]). Das ist wichtig, falls ein Term zwar eine hohe TF hat, aber in allen Dokumenten bzw. Textstellen sehr häufig vorkommen würde. Falls „Studentin“ im zuvor betrachteten Paragrafen

ebenfalls zweimal vorkommen würde, allerdings insgesamt viel häufiger auftritt als „Anmeldung“, wäre die Bedeutung dieses Terms geringer für die betrachtete Textstelle als „Anmeldung“. TF-IDF ist in der Lage, diesen Sachverhalt abzubilden.

$$\mathrm{IDF}(t, D) = \log \frac{N}{n_t}$$

Dabei gilt zusätzlich:

- D ist die Menge aller Dokumente (Korpus),
- $N = |D|$ ist die Gesamtanzahl der Dokumente im Korpus,
- n_t ist die Anzahl der Dokumente, in denen der Term t vorkommt.

3.3.2 Pretrained Language Model (PLM)

PLMs erzeugen Embeddings, die es ermöglichen, die semantische Bedeutung von Wörtern im Vektorraum-Modell darzustellen (Wang et al. [2023a]). Diese Embeddings können u.a. für Information Retrieval genutzt werden, indem typischerweise von Textpaaren Ähnlichkeitswerte berechnet werden, häufig basierend auf Skalarberechnungen wie der Kosinus-Ähnlichkeit (Vasileiou and Eberle [2024]). Besonders häufig werden Modelle, die auf BERT aufbauen, für Berechnungen der Textähnlichkeit eingesetzt (Vasileiou and Eberle [2024]). Denn im Gegensatz zu früheren Modellen wie GloVE (Pennington et al. [2014]) oder Word2Vec (Mikolov et al. [2013]) generiert BERT Embeddings, die die Bedeutung von Wörtern in Bezug auf ihre Verwendung im Kontext berücksichtigen (Gardazi et al. [2025]).

BERT (Bidirectional Encoder Representations from Transformers) wurde 2018 von Forschenden in Google AI Language entwickelt (Devlin et al. [2019]). Ein besonderes Merkmal von BERT ist die Bidirektionalität des Modells, die es ihm ermöglicht, die Bedeutung von Sätzen mit größerer Genauigkeit und Tiefe zu erfassen als Modelle, die den Kontext von Wörtern nur aus einer Richtung betrachten (Dejprapatsorn et al. [2025]). Diese Fähigkeit erlangt BERT durch sein Vortraining im Rahmen eines masked language model (MLM), bei dem zufällig ausgewählte Token verdeckt werden und das Modell trainiert wird, dieses Token vorherzusagen (Devlin et al. [2019]). Dieses Vorgehen unterscheidet sich von rekurrenten neuronalen Netzen (RNNs), die die Wörter normalerweise nacheinander verarbeiten, oder von autoregressiven Modellen wie GPT, die die zukünftigen Token intern maskieren (Devlin et al. [2019]). Dadurch kann das Modell eine bidirektionale Darstellung des Satzes erlernen (Devlin et al. [2019]). Zusätzlich zum MLM wird in den meisten auf BERT aufbauenden Modellen Next Sentence Prediction (NSP) eingesetzt: Bei dieser Technik verknüpfen die Modelle während des Vortrainings zwei maskierte Sätze, die manchmal im Originaltext nebeneinander standen, manchmal jedoch auch nicht (Devlin et al. [2019]). Dabei muss das Modell dann vorhersagen, ob die beiden Sätze im Originaltext aufeinander folgten oder nicht. Das Ziel ist es, durch die Anwendung von NSP die Leistung bei bestimmten Aufgaben, für die das Modell später eingesetzt werden soll und für die ein solches Wissen nützlich ist, zu verbessern (Liu et al. [2019]).

Konkret sollen folgende, auf BERT aufbauende Modelle eingesetzt werden[3]:

German BERT base German BERT (GBERT) base ist ein deutsches Sprachmodell und wurde 2020 veröffentlicht (Chan et al. [2020]). Es wurde mit einer größeren Menge an Trainingsdaten im Vergleich zu DBMDZ BERT, dem vorherigen besten deutschen BERT-Modell, trainiert: Hauptsächlich wurde mit dem Textkorpus OSCAR (145 GB Text, entspricht 89% der Trainingsdaten) sowie neben anderen Quellen auch mit OpenLegalData, trainiert (Chan et al. [2020]). Der OSCAR-Datensatz verwendet Texte, die aus dem Internet zusammengetragen wurden (Chan et al. [2020]). Im Gegensatz zu anderen BERT-Modellen wurde Whole Word Masking anstelle des bereits beschriebenen MLM verwendet: eine Technik, bei der das Modell das gesamte Wort finden muss, anstatt nur ein Token, das nur einen Wortteil repräsentieren könnte, verwendet (Cui et al. [2021]), um bessere Ergebnisse bei der Sprachmodellierung zu erzielen. Während des Trainings bewerteten (Chan et al. [2020]) die Modelle kontinuierlich, im Unterschied zu anderen Modellen, bei denen bereits vor dem Pretraining die Dauer des Trainings festgelegt wird.

Das Modell erzielte folgende Ergebnisse: GermEval18 Coarse: 78,17%, GermEval18 Fine: 50,90% und GermEval14: 87,98%. GermEval18 ist ein Twitter-Datensatz mit 5.000 deutschen Tweets, die als Missbrauch, Beleidigung, Obszönität und Sonstiges/Normal gekennzeichnet sind (Wiegand et al. [2018]).

BERT multilingual base model (cased) Multilingual BERT (ML BERT) ist ein mehrsprachiges Modell für 104 verschiedene Sprachen, darunter auch Deutsch, das auf Wikipedia-Daten trainiert wurde (Chan et al. [2020]). Es nutzt Masked Language Modeling (MLM) und Next Sentence Prediction (NSP), um kontextuelle Sprachrepräsentationen zu lernen (Devlin et al. [2019]. Es ist cased, somit unterscheidet es zwischen Groß- und Kleinschreibung.

German ELECTRA Base German ELECTRA (GELECTRA) wurde gemeinsam mit GBERT 2020 veröffentlicht (Chan et al. [2020]). GELECTRA basiert auf der ELECTRA Architektur, die 2020 von Clark et al. [2020] veröffentlicht wurde. Im Gegensatz zu den BERT basierten Modellen, die im MLM das Token, das durch das Modell im Training vorhergesagt werden soll, mit einer Maske ersetzt, wird hier das entsprechende Token durch ein anderes, ebenfalls plausibles ersetzt (Clark et al. [2020]). Das geschieht durch die Verwendung eines kleinen Generator-Networks und führt dazu, dass das Modell eine verbesserte Repräsentation des Kontext erlernt (Clark et al. [2020]). Das Modell zeigte eine etwas geringere Leistung als GBERT mit folgenden Werten: GermEval18 Coarse: 76,02%, GermEval18 Fine: 42,22% und GermEval14: 86,02%.

3.3.3 Large Language Model (LLM)

Anders als bei traditionellen Methoden oder auch Pretrained Language Models (PLMs), bei denen Vektorrepräsentationen mithilfe der Kosinus-Ähnlichkeit verglichen werden können, nutzt man typischerweise Prompting, um das Modell zum Re-Ranking möglicherweise relevanter Dokumente zu veranlassen (Zhuang et al. [2024]). Verwendet

[3]Der Fokus soll in dieser Arbeit auf den Modellen German BERT (GBERT) und multilingual BERT (ML BERT) liegen.

man LLMs ohne Fine-Tuning, um Dokumenten zu ordnen, lässt sich das Prompting in der Regel in drei Kategorien unterteilen: Pointwise, Pairwise und Listwise (Zhu et al. [2024]). Bei Pointwise-Methoden wird die Ähnlichkeit zwischen einer Anfrage und einem Dokument betrachtet, und die Antwort, die ein LLM liefert, ist häufig binärer Natur, wobei auch eine mehr gestufte Kategorisierung denkbar ist (Zhu et al. [2024]). Listwise-Methoden zielen darauf ab, eine Liste von Dokumenten zu sortieren (Zhu et al. [2024]). Der Nachteil dieses Vorgehens ist jedoch, dass LLMs in der Regel nur eine beschränkte Eingabe erlauben und ihre Performanz zu einem hohen Grad davon abhängt, wie die Liste, die eingegeben wird, sortiert ist (Sun et al. [2023]). Bei paarweisen Methoden erhalten LLMs eine Eingabe, die aus einer Anfrage und einem Dokumentenpaar besteht und werden angewiesen, die Kennung des Dokuments mit höherer Relevanz zu generieren (Sun et al. [2023]). Da im vorliegenden Kontext nur eine Textstelle aus der Liste der Textstellen relevant ist, bietet es sich an, eine Mischung aus diesen Methoden zu verwenden: Dem LLM wird eine Liste präsentiert, aus der die Textstelle, die die höchste Relevanz hat, ausgegeben werden soll.

Um ein LLM auszuwählen, wurden verschiedene frei verfügbare LLMs getestet, wobei Mistral Large vielversprechende Ergebnisse lieferte. Mistral Large ist ein Open-Source-LLM, das für seine robuste Leistung bei verschiedenen Aufgaben des Verständnisses natürlicher Sprache bekannt ist (McDonald et al. [2024]). Die Architektur von Mistral ermöglicht es ihm, komplexe Themen zu erfassen und Bedeutungen aus dem Kontext abzuleiten (Tsai et al. [2024]). Seine Anwendung geht über die bloße Textgenerierung hinaus und erstreckt sich auf Bereiche, die ein tiefes Verständnis und das Zusammenfassen von Informationen erfordern (Tsai et al. [2024]). Mistral Large 2411 wurde von Mistral AI im November 2024 veröffentlicht und ist eine weiterentwickelte Version der Vorgängervariante Mistral Large 2407. Mistral Large gehört mit etwa 123 Milliarden Parametern zur Klasse der großen Sprachmodelle (LLMs) (Mistral AI [2024]). Das Modell bietet eine maximale Kontextlänge von 128.000 Tokens, wodurch es in der Lage ist, sehr lange Texte zu verarbeiten (Mistral AI [2024]). Es unterstützt viele Sprachen, darunter auch Deutsch, und wurde darauf trainiert, zu erkennen, wenn es keine Lösungen finden kann oder nicht über ausreichende Informationen verfügt, um eine sichere Antwort zu geben (Mistral AI [2024]).

3.4 Evaluierung

3.4.1 Definition der Bewertungsgrundlage

Letztendlich existieren IR-Systeme mit dem Ziel eine Interaktion zwischen den Produzierenden von Informationen und den Nutzenden in Form einer sozialen Interaktion zu ermöglichen (Saracevic [1995]) und den Zugang zu Informationen zu erleichtern (Arora et al. [2016]). Um diese Komponente zu berücksichtigen, bedarf es neben einer systemischen Evaluation daher auch einer nutzenden-zentrierten Evaluation (Saracevic [1995]). Die Evaluation von IR-Systemen ist häufig auf dem Kriterium der Relevanz aufgebaut (Saracevic [1995]). Allen et al. [1955] sieht Relevanz sogar als das einzige Kriterium an, nach dem ein IR-System evaluiert werden kann. Relevanz ist jedoch subjektiv, hängt von der Erwartung der nutzenden Person ab und lässt sich schlecht messen (Arora et al. [2016]). Daher sollte die Bewertung, inwiefern die Ergebnisse eines IR-Systems für die suchende Person relevant sind, durch diese selbst geschehen, je nach ihren Fähigkeiten und Bedürfnissen (Hoffmann et al. [2011]). In der vorliegenden Arbeit wird davon ausgegangen, dass die Beurteilung nach Relevanz der gefundenen Textstellen für alle Nutzenden sehr ähnlich aussehen würde.

Denn entweder wird durch die gegebene Textstelle die TAQ beantwortet oder nicht. Daher wurde sich dazu entschieden, die Textstellen nach eigenem Ermessen nach Relevanz (binär) einzuteilen, um die Performanz der Ansätze bewerten zu können.

3.4.2 Umfrage 2 und Evaluationsmetrik

Ziel der Umfrage 2

Im vorliegenden IR-System befindet sich die Antwort jeder TAQ in genau einer Textstelle. Die Anwendung von Metriken, die darauf ausgerichtet sind, dass mehrere Textstellen bzw. Dokumente relevant sind, bietet sich demnach nicht an. Viele der bestehenden Metriken bauen jedoch auf dieser Annahme auf (Bama et al. [2018]). Dazu gehören u.A. weit verbreitete und etablierte Metriken, die auf Recall und Precision aufbauen (Arora et al. [2016]): Im gegebenen Kontext ist ihre Aussagekraft beschränkt, da Recall@k 100% wäre, sobald die eine relevante Textstelle gefunden wurde, und ansonsten 0%. Recall@k ist somit nicht differenziert genug, um eine Evaluation durchzuführen und hängt maßgeblich von der Wahl des k ab. Zusätzlich spricht gegen die Verwendung von Recall die Tatsache, dass es in keinem Zusammenhang mit der Zufriedenheit der Nutzenden steht (Moffat and Zobel [2008]). Analoges gilt für die Verwendung von Precision.

Außerdem gibt es einen klar abgrenzbaren Suchraum mit einer zum Vergleich von Web-Suchen eher kleineren Anzahl an möglichen Treffern. Die bestehenden Metriken sind allerdings häufig im Kontext der Internetsuche, bei der in großen Sammlungen gesucht wird, entwickelt worden (Bama et al. [2018]). Bei diesen Metriken kann die Gesamtanzahl somit beim Vergleich zwischen verschiedenen Systemen vernachlässigt werden, da die Suchräume generell sehr groß sind.

Um diese Kriterien bestmöglich zu erfüllen und insbesondere um die Passung zur Zielgruppe beurteilen zu können, soll mit Hilfe der Ergebnisse aus Umfrage 2 eine eigene Metrik aufgebaut werden. Das wichtigste Maß für die nutzenden-zentrierte Evaluation eines IR-Systems ist die Zufriedenheit der Nutzenden, die maßgeblich von der Antwortzeit und der Relevanz der Ergebnisse sowie von der Benutzeroberfläche abhängt – einschließlich Design, Klarheit, Präzision, Reaktionsgeschwindigkeit und inhaltlicher Relevanz (Arora et al. [2016]). Da das System nicht im Sinne eines vollständigen IR-Systems implementiert wird, braucht es einer anderen Lösung, um eine nutzenden-zentrierte Evaluation durchzuführen. Um herauszufinden, mit welcher Systemperformanz abseits der Gestaltung einer möglichen Benutzeroberfläche die Nutzenden zufrieden wären, eignet sich ein Fragebogen mit einem fiktiven Szenario und nur angedeuteter Benutzeroberfläche.

Um einen solchen Fragebogen zu entwerfen, benötigt es ein Modell des Suchverhaltens von den Nutzenden, denn eine Metrik zum Evaluieren eines IR-Systems hängt von diesem zugrundeliegenden Modell ab (Chapelle et al. [2009]). Analog zur Arbeit von Moffat and Zobel [2008] wird in diesem Kontext davon ausgegangen, dass Nutzende beim ersten Dokument starten, zu prüfen, ob es zur Anfrage passt, und dann mit dem nächst relevant bewerteten Dokument fortfahren. Im gegebenen Kontext ist maximal eine Textstelle relevant, das bedeutet, dass die Suchdauer (aufgefasst als Anzahl der Dokumente, die aussortiert werden müssen für jedes relevante Dokument (Moffat and Zobel [2008]) ergibt sich hier aus der Stelle, in der die relevante Textstelle gefunden wurde. Wurde die relevante Textstelle zum Beispiel an 4. Stelle gefunden, ergibt sich eine kürzere Suchdauer als in dem Fall, dass die relevante Textstelle an 10. Stelle gefunden wurde.

Somit stellt sich die Frage, ob Studierende (bzw. Suchende) im gegebenen Kontext eher „geduldig“ oder „ungeduldig“ sind, mit anderen Worten, wie viele Textstellen sie bereit sind zu betrachten. Genau das soll mit der Umfrage 2 herausgefunden werden.

Entwicklung der eigenen Metrik

Betrachtet man die Items, die nach der Stelle, an der eine Textstelle maximal stehen darf, damit die Teilnehmenden das IR-System als noch nützlich empfinden und weiterempfehlen würden, ergibt sich ein Mittelwert von $M = 4{,}40$ ($SD = 2{,}65$). Für die Bewertung als weder zufrieden noch unzufrieden ergab sich entsprechend $M = 6{,}49$ ($SD = 4{,}36$) und für unzufrieden und nutzlos $M = 7{,}78$ ($SD = 5{,}47$). Die genauen Kennwerte pro Item können im Anhang gefunden werden (siehe Tabelle 22).

Diese empirischen Werte sollen sich in der Bewertung des Ranges, in dem die relevante Textstelle gefunden wurde, widerspiegeln. Als Orientierung soll hier die statistische Kennzahl Cohen's d, ein Maß zur Quantifizierung eines empirischen Effekts im Rahmen von statistischen Tests, dienen (Cohen [2013]). Nach Rice and Harris [2005] hat Cohen die Werte 0,2 (kleiner Effekt), 0,5 (mittlerer Effekt) und 0,8 (großer Effekt) jeweils mit Beispielen aus der Realität in Verbindung gebracht.

Dieses Vorgehen wurde hier adaptiert: Die Stufen, die Cohen verwendet (0,2, 0,5 und 0,8) sollen eine schlechte, mittlere und gute Performanz des Ansatzes widerspiegeln. Dies geschah ebenfalls in der Verankerung der Realität, in diesem Fall durch die Ergebnisse aus der Umfrage 2.

Insgesamt ergaben sich folgende Anforderungen an die Funktion, die die Metrik verwenden soll:

- streng monoton fallend in $(1, \infty)$
- Wendepunkt bei $x \approx 6{,}49$, $y \approx 0{,}50$ (mittlere Performanz)
- Schnitt mit $x \approx 4{,}40$, $y \approx 0{,}80$ (gute Performanz) und $x \approx 7{,}78$, $y \approx 0{,}20$ (schlechte Performanz)
- positiv in $(1, \infty)$

Eine Funktion, die diese Anforderungen erfüllt, lautet:

$$f(x) = \frac{1}{0.99 + e^{0.8(x-6.5)}}$$

In Abbildung 2 ist sowohl diese Funktion als auch der Reciprocal Rank in Abhängigkeit der verschiedenen Ränge (x-Werte) dargestellt.

3.4.3 Evaluation der Anwendung von Mistral Large

Da die Antworten von Mistral Large in Textform vorliegen, sollen sie händisch ausgewertet werden. Um möglichst aussagekräftige Ergebnisse zu erhalten, sollen mehrere Durchläufe ausgewertet werden, und für die weitere Interpretation die Antworten verwendet werden, die in den Durchläufen weitgehend übereinstimmten.

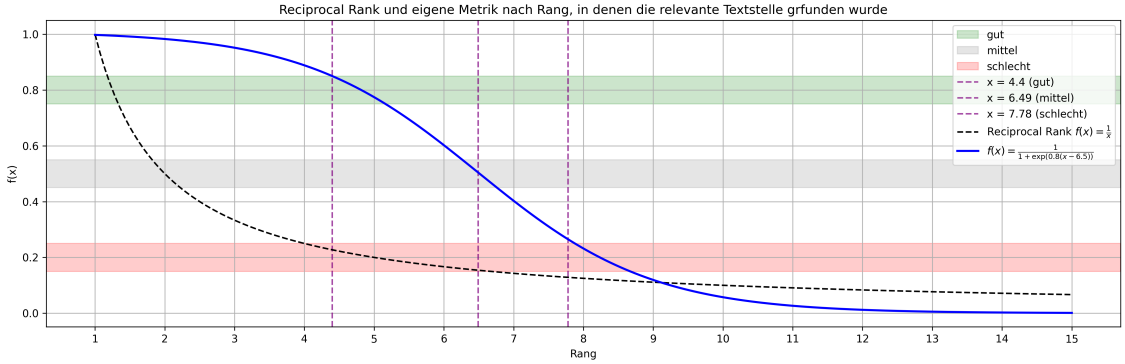

Abbildung 2: *Vergleichende Darstellung der Reciprocal Rank und der eigenen Metrik nach Rang der relevanten Textstelle.*

3.4.4 Einteilung der TAQ nach Art der Frage und Thema

Um zu prüfen, ob es Kriterien in den TAQ gibt, die begründen, warum bestimmte Ansätze gut oder schlecht funktionieren, sollen die TAQ zusätzlich in Gruppen eingeteilt werden. Zum einen ausgehend von der Art der Frage, die die TAQ stellt, zum anderen nach dem Thema, zu dem die TAQ gehört.

Für die Einteilung nach Art der Frage wird die Typisierung nach Kolomiyets and Moens [2011] in *Faktfrage*, *Listenfragen*, *Definitionsfragen* und *komplexe Fragen* verwendet:

- Faktfragen sind Fragen, die nach einer einfachen Tatsache fragen und in wenigen Wörtern beantwortet werden können, wie zum Beispiel: *Wie weit ist es von der Erde zum Mars?* (Heie et al. [2012]).
- Bei Listenfragen wird als Antwort eine Menge von Einheiten erwartet, die bestimmte Kriterien erfüllen, ein gutes Beispiel hierfür ist: *Wann hat Brasilien die Fußball-Weltmeisterschaften gewonnen?* (Heie et al. [2012]).
- Definitionsfragen erwarten als Antwort eine Zusammenfassung oder einen kurzen Abriss, wie etwa die Frage: *Wie funktioniert die Mitose einer Zelle?* (Neves and Leser [2015]).
- Im Gegensatz dazu beziehen sich komplexe Fragen auf Informationen in einem bestimmten Kontext (Soares and Parreiras [2020]).

Allerdings zeigt sich die Unterscheidung zwischen *Definitionsfragen* und *komplexe Fragen* in der Praxis nicht eindeutig umsetzbar, daher werden diese beiden Kategorien zusammengefasst als *Erweiterte Fragen*.

3.4.5 Methoden aus der Statistik

Um die Ergebnisse der Ansätze zu vergleichen, soll auf Methoden aus der Inferenzstatistik zurückgegriffen werden. Ziel hierbei ist es abzuschätzen, mit welcher Wahrscheinlichkeit bestimmte Ergebnisse auftreten. Dabei wurde ein Between-Subjects-Design mit $N = 20$ pro Gruppe angenommen, da es sich um 20 TAQs handelt [4]. Nicht-parametrische Tests setzen weniger voraus als ihre parametrischen Entsprechungen, zum Beispiel Annahmen

[4] Auf Subparagrafenebene lediglich $N = 14$, daher wurde für die Analysen häufig auf die Paragrafenebene zurückgegriffen.

über die Verteilung, aus der gezogen wird, und werden daher in der Praxis häufig verwendet (Hoskin [2012]). Der p-Wert gibt an, mit welcher Wahrscheinlichkeit, unter der Annahme, dass die Nullhypothese gilt, die beobachteten Ergebnisse aus einer Verteilung stammen, die für die Alternativhypothese spricht (Goodman [2008]). Die Nullhypothese geht davon aus, dass zwischen zwei Gruppen kein Unterschied vorliegt, während die Alternativhypothese das Gegenteil besagt (Hilgers et al. [2018]). Wie üblich soll auch in dieser Arbeit ein p-Wert von $p = 0{,}05$ verwendet werden.

Kapitel 4

Umsetzung

4.1 Empirischer Teil

Die Umsetzung der Umfrage 1 und Umfrage 2 geschah mithilfe der Plattform `www.umfrageonline.com`. Die Rekrutierung der Teilnehmenden geschah über Bekannte aus dem persönlichen Umfeld. Die Auswertung geschah mithilfe von Microsoft Excel und Python [1]. Um möglichst viele Personen ansprechen zu können, wurden beide Fragebögen in zwei Varianten angeboten: Eine deutsche und eine englische Version mit gleichem Inhalt.

4.1.1 Umfrage 1

Ziel der Umfrage 1 („Mein Studium, meine Ordnung – meine Ideen“) war es, für Studierende relevante typische Fragen herauszufinden. Zusätzlich wurde erhoben, wie wichtig bestimmte Themen für Studierende waren und wie viel ihrer Studien- und Fachprüfungsordnung sie gelesen haben. Der Fragebogen richtete sich ausschließlich an Studierende der Universität Bamberg. Bei der englischen Version wurden bestimmte Fachausdrücke (z.B. „Allgemeine Prüfungsordnung“ sowohl auf Deutsch als auch auf Englisch beschrieben. Um mögliche Unklarheiten und Verwechslungen zu vermeiden, wurden Begriffe wie „Studien- und Fachprüfungsordnung“ sowie „Fachsemester“ erklärt. In der Einleitung wurde zusätzlich ein Link zu den Studienordnungen bereitgestellt. Die Teilnahme war anonym und freiwillig.

Die Items wurden anhand der Fragestellung entwickelt: Neben allgemeinen Fragen zu der Person (Alter, Studiengang, Fachsemester) wurden insbesondere Items genutzt, die typische Fragen Studierender bezüglich ihrer Studien- und Fachprüfungsordnung herausfinden sollten. Dazu wurde direkt nach Fragen gefragt, aber auch nach Themen möglicher oder eventueller Konflikte und Beratungsgespräche. Hierfür wurden Freitextfelder eingesetzt. Um zu erfahren, wie hoch der Anteil der gelesenen Studien- und Fachprüfungsordnung war, wurde ein Schieberegler eingesetzt. Eine Gewichtung verschiedener Themen nach Wichtigkeit geschah über eine fünfstufige Likert-Skala, somit erhielten die Teilnehmenden die Möglichkeit, neutral zu antworten. Die Antwortmöglichkeiten waren inhaltlich symmetrisch verankert von „sehr irrelevant“, „etwas relevant“, „neutral“, „etwas wichtig“ und „sehr wichtig“. Die abgefragten Themen umfassten sowohl Themen, die die Studien- und Fachprüfungsordnung beinhalten, als auch weitere Themen, wie z.B.

[1] Das Material kann im GitLab gefunden werden.

„Auslandsaufenthalte“. Die Themenabfrage geschah zum Ende des Fragebogens, das letzte Item war ein Textfeld für weitere Anregungen und Feedback seitens der Teilnehmenden.

4.1.2 Umfrage 2

Ziel der Umfrage 2 war es herauszufinden, mit welcher Performanz eines IR-Systems die Suchenden zufrieden wären, sie es als nützlich empfinden würden und weiterempfehlen würden. Es gab keine Voraussetzungen, um an dem Fragebogen teilzunehmen. Zusätzlich wurde erfasst, ob die teilnehmende Person sich momentan in einem Studium befand, falls nein, ob sie zuvor einmal studiert hat. Mit Hilfe eines Szenarios, in dem ein Tool vorgestellt wird, das der Person beim Finden einer Frage zu Studienordnung oder zum Modulhandbuch helfen soll, indem es die relevante Textstelle vorschlägt. Im Anhang ist die entsprechende Seite der Umfrage zu finden (siehe Abbildung 14). Die Zufriedenheit wurde mit drei Items erfasst, die jeweils nach einer Zahl fragen: „Bitte gebe die maximale Stelle an, an der die relevante Stelle stehen darf, damit du mit dem Tool zufrieden bist:“, „Bitte gebe die durchschnittliche Stelle an, an der die relevante Stelle stehen muss, damit du mit dem Tool weder zufrieden noch unzufrieden bist:“ und „Bitte gebe die erste Stelle an, an der die relevante Stelle stehen muss, damit du mit dem Tool unzufrieden bist:“. Die Erfassung der Nützlichkeit erfolgte analog, für die Frage, mit welchem Rang die Teilnehmenden das Tool weiterempfehlen würden, wurde ebenfalls ein analoges Item verwendet. Zum Schluss hatten die Teilnehmenden die Möglichkeit, in einem offenen Textfeld Feedback oder weitere Anregungen und Gedanken mitzuteilen.

4.2 Datengrundlage und -aufbereitung

Bei den Methoden, bei denen mit Hilfe von Ähnlichkeitsmaßen die Textstellen nach Relevanz für die TAQs sortiert wurden (betraf TF-IDF und auf BERT aufbauende Modelle), wurden die Ergebnisse ebenfalls im .json-Format gespeichert. Die so entstandenen Dateien können im GitLab gefunden werden.

Den Textkorpus bildeten die Studien- und Fachprüfungsordnungen (StuFPO) der einzelnen Studienfächer, sowie die Allgemeinen Prüfungsordnungen (APO) der entsprechenden Fakultäten[2]. Diese wurden von der Homepage (`https://www.uni-bamberg.de/abt-studium/aufgaben/pruefungs-studienordnungen/`[3]) der Universität Bamberg heruntergeladen. Danach wurde nicht mehr auf Aktualität der Ordnungen geprüft. Diese wurden entsprechend ihrer Nummerierung zusammengefasst und im .json-Format gespeichert. Dabei wurde jeder Paragraph einzeln mit folgender Unterteilung gespeichert: Die Nummer des Paragraphen, der Titel des Paragraphen, sofern vorhanden die Unterparagraphen sowie der komplette Text samt dem Titel. Diese Unterteilung wurde vorgenommen, um die verwendeten Methoden an den verschiedenen Textebenen eines Paragraphen testen zu können. Die Unterteilung erfolgte unter Verwendung von Regulären Ausdrücken. Der Anhang wurde wie ein Paragraf verwendet und manuell, lediglich für die verwendete Studienordnung, eingeteilt. Der entsprechende Code und Datensatz ist im GitLab zu finden.

Um möglichst viele TAQs herauszufinden, wurden die TAQs anhand der Fragen, die die Studierenden in Umfrage 1 nannten, aber auch anhand ihrer Konflikte bzw. Konflik-

[2]StuFPO und APO werden im Folgenden der Einfachheit wegen als Studienordnung bezeichnet.
[3]zuletzt abgerufen am 15.11.2024

te, die sie sich vorstellen konnten, zu haben, abgeleitet. In einem weiteren Schritt wurde händisch bewertet, in welcher Textstelle die die TAQ beantwortende Information zu finden ist. Dies geschah für die Studienordnung des Masterstudiengangs CitH, indem der Paragraf, der die TAQ beantwortet, bzw. falls vorhanden der Unterparagraf, händisch in einer Excel-Tabelle hinterlegt wurde. Bei 14 der 20 TAQs waren Unterparagrafen vorhanden. Außerdem wurde jeder TAQ eine ID zugewiesen (beginnend mit taq_0 bis taq_19, siehe Tabelle 21 im Anhang für eine Übersicht). Für die weitere Verarbeitung wurden die TAQs inhaltlich in vier Gruppen händisch eingeteilt, diese waren „Inhalt des Studiums“, „Dauer des Studiums“, „Abschlussarbeit“, „Prüfungen“ und je nach Frageart in drei Gruppen: „Faktfragen“, „Listenfragen“ und „Erweiterte Fragen“. Sowohl die TAQs, als auch die Studienordnung des Masterstudiengangs CitH wurden unter Verwendung des Übersetzers „DeepL[4]“ übersetzt. Um die Studienordnungen und die TAQs ins Englische zu übersetzen wurde DeepL [5], das für seine Genauigkeit bekannt ist und hohe Scores sowohl in BLEU als auch in der Evaluation durch Menschen erreicht, genutzt (Kamaluddin et al. [2024]).

4.3 Methoden

Um die verschiedenen Ansätze testen zu können, wurde sich auf die Studienordnung des Masterstudiengangs „Computing in the Humanities M. Sc“ fokussiert. Bei der Berechnung der Metriken für TF-IDF und den auf BERT aufbauenden Modellen wurde darauf geachtet, dass, sobald die Cosinus-Ähnlichkeit der tatsächlich relevanten Textstelle null war, das Ergebnis der Metrik auch auf null gesetzt wurde. Dieses Vorgehen verhindert zufällige Ergebnisse, insbesondere bei der Evaluation von TF-IDF, da hier besonders häufig die Cosinus-Ähnlichkeiten null waren. Die auf Umfrage 2 aufbauende Metrik wurde entsprechend der Formel (vergleiche Abschnitt 3.4.2) implementiert, der entsprechende Code ist im GitLab zu finden.

4.3.1 TF-IDF

Die Texte wurden von überflüssigen Leerzeichen und Sonderzeichen bereinigt. Um die Stoppwörter zu entfernen, wurde auf die Stoppwörter zurückgegriffen, die im Modul nltk.corpus.stopwords bereitgestellt werden. Für die deutschen Texte wurde der „GermanStemmer“ aus dem Modul `nltk.stem.snowball` verwendet, für die englischen Übersetzungen die entsprechende englische Version. Es wurde der TfidfVectorizer von scikit-learn verwendet, der aus den Texten eine sogenannte TF-IDF-Matrix erstellt. Anschließend wurden die Kosinus-Ähnlichkeiten unter Verwendung des Moduls sklearn.metrics.pairwise berechnet. Der entsprechende Code ist im GitLab zu finden.

4.3.2 BERT-basierte Modelle

Zur Berechnung semantischer Ähnlichkeiten auf Satzebene wurden vortrainierte Sprachmodelle (PLMs) verwendet, die auf BERT (Bidirectional Encoder Representations from Transformers) basieren. Zum Einsatz kommt das Modul `SentenceTransformer` aus

[4]im Probeabonnement

[5]https://www.deepl.com/de/translator, genutzt als Probeabo

der Bibliothek `sentence_transformers`, das auf den `transformers`-Modellen von Hugging Face aufbaut. Anschließend wurden, analog zum TF-IDF-Ansatz, die Kosinus-Ähnlichkeiten zwischen diesen Embeddings berechnet. Dafür wurde ebenfalls das Modul `sklearn.metrics.pairwise` verwendet. Der entsprechende Code ist im GitLab zu finden.

4.3.3 Mistral Large

Für die Anwendung eines Large Language Models (LLM) wurde das Modell Mistral Large (Version 2411) verwendet. Die Anfragen (TAQs) wurden per API an das Modell übergeben. Die vom Modell zurückgegebenen Textstellen wurden gemeinsam mit den Eingabetexten in einem `pandas.DataFrame` gespeichert. Die Bewertung der Ergebnisse des LLMs konnte aufgrund des Formats des Outputs nicht automatisiert erfolgen, somit erfolgte die Bewertung händisch.
Der Prompt lautete:

Die Frage ist: {taq}
Suche aus den Texten den Text, der die Frage am besten beantwortet.
Die Texte sind:
{texts}
Gebe die Antwort nicht selbst.
Gebe einfach den Text zurück, der die Antwort enthält.

Sofern eine Textstelle korrekt gefunden wurde, wurde das Ergebnis als korrekt gewertet, solange der Wortlaut mit dem gegebenen Text übereinstimmt. Wurde jedoch die korrekte Antwort gegeben, wurde das Ergebnis als falsch gewertet, da nach der Textstelle und nicht der direkten Beantwortung gesucht wurde. Werden bei Subparagrafen mehrere Subparagrafen eines Paragrafen zurückgegeben, so wurde zumindest der Paragraf bewertet. Werden mehrere Subparagrafen verschiedener Paragrafen zurückgegeben, wird das Ergebnis als falsch gewertet.

4.4 Query Expansion

Um zu testen, inwiefern Query expansion eingesetzt werden kann, um die Ergebnisse der Verwendung der ins Englische übersetzten TAQs und Studienordnung zu verbessern, wurde mit folgendem Prompt an ChatGPT (Version GPT-4o) von OpenAI (über die WebApp https://chat.openai.com), mit jeweils einer TAQ, gearbeitet[6]:

„I have a set of questions related to student regulatory content. I want to expand these questions to generate a broader set of semantically similar queries that may include synonyms, rephrased versions, and closely related topics. These expanded queries will be used in a TF-IDF-based information retrieval system, so diversity and coverage are important, but they should still remain topically relevant to the original intent. Please provide 5–10 expanded versions of the following question, ensuring that the language varies but the core

[6]Die so gefundenen Anfragen können im GitLab gefunden werden.

meaning is preserved. If relevant, include alternative phrasing used by students, faculty, or administrative staff. This is the question: ... “

4.5 Vortests

Für alle verwendeten Ansätze wurden Vortests mit Texten und Anfragen, die sehr einfach zuzuordnen sind, durchgeführt, um sowohl explorativ verschiedene Modelle, insbesondere PLMs und LLMs, zu testen, als auch um zu überprüfen, ob die verwendeten Ansätze richtig implementiert sind und plausible Ergebnisse liefern. Das Vorgehen war genauso wie bei dem beschriebenen Vorgehen zur Datengrundlage. Als TAQs wurden jedoch einfach die ersten Sätze des verwendeten Beispieltextes benutzt. Für die vorliegende Arbeit wurden ein Paragraph aus der verwendeten Studienordnung, ein Wikipedia-Artikel über einen Künstler sowie ein Textausschnitt aus „Die Leiden des jungen Werthers“ von Johann Wolfgang von Goethe herangezogen.

Kapitel 5

Ergebnisse

5.1 Ergebnisse aus den Vorbereitungen

5.1.1 Ergebnisse der Umfrage 1 und Beschreibung der TAQs

Durch den Fragebogen 1 konnten insgesamt 20 TAQs identifiziert werden. Die vollständige Liste aller TAQs ist im Anhang zu finden (siehe Tabelle 21). Die thematische Verteilung ist in Tabelle 1 dargestellt, und die Einteilung der Fragen nach Art der Frage in Tabelle 2. Zwar handelt es sich größtenteils um „Faktfragen“, wie beispielsweise „*Kann ich Pflichtmodule durch andere Module ersetzen?*“, jedoch weniger als erwartet. Mit jeweils 6 Fragen gab es gleich viele „Listenfragen“ wie „erweiterte Fragen“. Ein Beispiel für eine „Listenfrage“ ist „*Was sind die verschiedenen Prüfungstypen?*“ und für eine „Erweiterte Frage“ „*Was hat es mit der Studienfortschrittskontrolle auf sich?*“. Als besonders wichtig schätzten die Teilnehmenden die Themen „Vorgaben zu Abschlussarbeiten“, „Aufbau des Studiums“und „Module und Modulkataloge„ein. Am wenigsten relevant wurde das Thema „Auslandsaufenthalte“empfunden. Insgesamt zeigten sich jedoch keine großen Unterschiede in der Bewertung der Themen. Wie erwartet wurden insbesondere TAQs zu Themengruppen gefunden, die von den Studierenden als besonders wichtig eingeschätzt wurden. Die vollständigen Ergebnisse können im Anhang gefunden werden.

Der Anteil der gelesenen Studienordnung ist mit durchschnittlich 40% überraschend hoch. Es kann sein, dass für die Teilnehmenden der Unterschied zwischen dem Modulhandbuch und der Studienordnung unklar war. Die Verteilung der Ergebnisse kann vermuten lassen, dass es zwei Gruppen Studierender gibt: Solche, die einen großen Anteil ihrer Studienordnung gelesen haben, und solche, die maximal die Hälfte gelesen haben (vergleiche Abbildung 3).

Tabelle 1: *Verteilung der TAQS auf Themengruppen*

Themengruppe	Anzahl
Inhalt des Studiums	7
Dauer des Studiums	5
Abschlussarbeit	5
Prüfungen	3

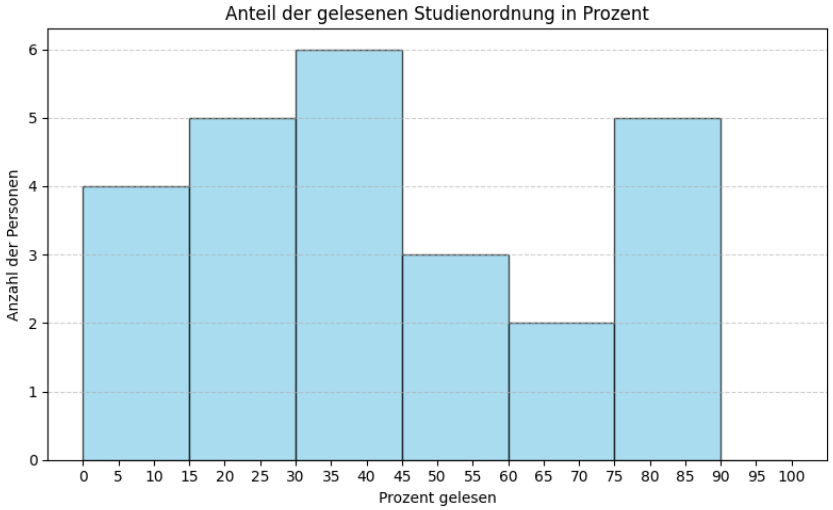

Abbildung 3: *Anteil der gelesenen Studienordnung in Prozent*

Tabelle 2: *Verteilung der TAQS nach Art der Frage*

Fragenart	Anzahl
Faktfrage	8
Listenfrage	6
Erweiterte Frage	6

5.1.2 Ergebnisse der Umfrage 2

In der zweiten Umfrage nahmen insgesamt 64 Personen teil, acht davon nutzen die Version in englischer Sprache. 25 Personen befanden sich zum Zeitpunkt der Umfrage in einem Studium, 29 Personen befanden sich nicht in einem Studium und hatten zuvor studiert, und 9 Personen haben nie studiert. Es handelte sich somit um eine eher akademisch geprägte Stichprobe.

Sowohl in der Bewertung der Zufriedenheit mit dem Tool, als auch in der Bewertung der Nützlichkeit des Tools und in der Frage, mit welchem Ergebnis die Teilnehmenden das Tool weiterempfehlen würden, häuften sich die Bewertungen in bestimmten Werten: Wie in Abbildung 4 zu erkennen, würden die Hälfte der Teilnehmenden mit dem Tool zufrieden sein, falls die relevante Textstelle maximal an 3. Stelle gefunden würde. Als nützlich und weiterempfehlen würden die Hälfte der Personen es, wenn die relevante Textstelle noch an 4. Stelle gefunden worden wäre. Wie erhofft, zeigt sich durch die Häufung in der Angabe der Ränge ein eindeutig interpretierbares Bild: Besonders im Antwortverhalten zu den Items, die nach einer positiven Eigenschaft des Systems (Zufriedenheit, Nützlichkeit und Weiterempfehlung) fragten, zeigten die Antworten eine geringe Varianz. Die größte Varianz zeigte sich in der Frage nach der Nutzlosigkeit. Die genauen Ergebnisse können im Anhang gefunden werden Tabelle 22. Diese Ergebnisse legen somit eine Grundlage, um die Ansätze anhand der Zielgruppe zu evaluieren.

Um zu prüfen, ob es signifikante Gruppenunterschiede zwischen den Personen, die

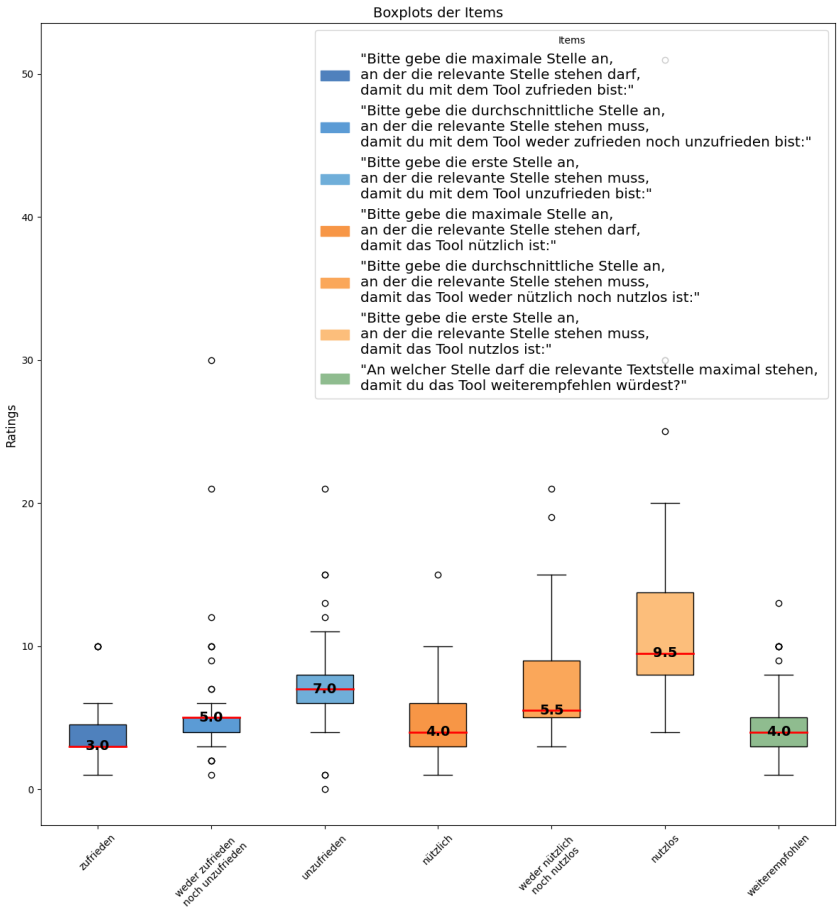

Abbildung 4: *Items und Antwortverhalten der Teilnehmenden in Umfrage 2*

zum Zeitpunkt der Umfrage studieren, und denjenigen, die zum Zeitpunkt der Umfrage nicht (mehr) studieren, gibt, wurde für jedes Item ein zweiseitiger T-Test durchgeführt. In keinem der Items zeigte sich ein signifikanter Gruppenunterschied. Dies spricht dafür, dass die Ergebnisse aus der Umfrage über eine studentische Zielgruppe hinaus generalisierbar sein könnten. Leider war die Teilnehmerzahl derjenigen, die nie studiert hatten, zu klein, um auf Unterschiede in dem Antwortverhalten zu testen.

Als Freitextrückmeldung kamen drei Anmerkungen: Zum einen, dass es schwierig sei, sich vorzustellen, mit welchem Rang man zufrieden wäre; eine Person würde den Ergebnissen nicht vertrauen und die Ergebnisse nochmal selbst überprüfen durch eine eigene Suche und letztendlich die Frage, warum das System existiert und man nicht einfach über „STRG plus F“ nach der relevanten Textstelle suchen würde.

5.1.3 Beschreibung der relevanten Textstellen

Wie in Tabelle 3 zu erkennen ist, ergaben sich, nachdem die Studienordnung in Paragrafen, Subparagrafen und Titel eingeteilt wurde, samt Anhang je 43 Paragrafen und Titel und 124 Subparagrafen. Die Mittelwerte der Anzahl der Wörter lagen bei $M = 257{,}81$ für Paragrafen, $M = 60{,}11$ für Subparagrafen und $M = 3{,}26$ für Titel.

Tabelle 3: *Deskriptive Kennzahlen der Wortanzahl der Studienordnung*

Textebene	**N**	**Min**	**Max**	**M**	**Median**	**SD**
Paragraf	43	20	1556	257,81	166	342,07
Subparagraf	124	6	341	60,11	44	53,25
Titel	43	1	10	3,26	3	2,26

Die Tabelle 4 zeigt den Vergleich zwischen der deutschen, originalen Studienordnung und der englischen bezüglich der Anzahl der Wörter und der einzigartigen Wörter im Vergleich. Ebenfalls ist zu erkennen, dass die deutschen Wörter durchschnittlich fast ein Viertel länger sind als die englischen. Das liegt vermutlich an den zusammengesetzten Nomen in der deutschen Sprache, die in zwei Wörter im Englischen übersetzt werden (wie z.B. „Masterarbeit“, das zu „master's thesis“ übersetzt wird). Des Weiteren sieht man, dass die deutsche Version einen fast doppelt so großen Anteil an einzigartigen Wörtern aufweist im Vergleich zur englischen Version. Daraus lässt sich schließen, dass in der deutschen Version die Begriffe, vermutlich insbesondere die zusammengesetzten Nomen, einen spezifischeren Gebrauch aufweisen als die Begriffe in der englischen Version, die ihre Bedeutung maßgeblich durch die umstehenden Begriffe erhalten. So enthält der Begriff „Masterarbeit“ mehr Informationen als der Begriff „thesis“, der erst durch seinen Kontext (z.B. „master's“) denselben Informationsgehalt hat.

Diese unterschiedlichen Eigenschaften der deutschen und der englischen Version der Studienordnung könnten einen Einfluss auf die Performanz der Methoden, insbesondere auf TF-IDF, haben. So führen viele spezifische Begriffe zu einer hohen Data Sparsity. Bei der Anwendung von TF-IDF entstehen somit Vektoren mit höheren Dimensionen auf der originalen Studienordnung im Vergleich zur englischen Übersetzung. In der Folge könnte die Kosinus-Ähnlichkeit in ihrer Fähigkeit, aussagekräftig zwischen ähnlichen und unähnlichen Objekten zu unterscheiden, verlieren (v.g.l. „Curse of Dimensionality“).

Tabelle 4: *Vergleich der Anzahl der Wörter und Wortlänge der ins Englische übersetzten und der deutschen Studienordnung*

	Englisch	**Deutsch**
Anzahl der Wörter	7181	6083
Anzahl der einzigartigen Wörter	887	1403
Durchschnittliche Wortlänge (in Zeichen)	5,81	8,03
Einzigartige Wörter / Anzahl der Wörter	0,124	0,231

5.2 Ergebnisse der Anwendung der Ansätze

Die Ergebnisse der Analyse werden anhand ausgewählter Fragestellungen dargestellt. Zur Bewertung der Anwendung von TF-IDF und den BERT-basierten Modellen soll die eigene, auf Umfrage 2 basierende Metrik verwendet werden.

5.2.1 Ergebnisse der Anwendung von TF-IDF

1. Wie gut funktioniert die Anwendung von TF-IDF im gegebenen Kontext?

2. Inwiefern beeinflusst Textlänge die Performanz von TF-IDF beim Ranking?
3. Inwiefern beeinflusst die Art der Frage die Performanz von TF-IDF beim Ranking?
4. Inwiefern verbessert die Übersetzung der Texte ins Englische die Performanz und warum?
5. Welchen Einfluss hat Pre-Processing auf die Performanz von TF-IDF?
6. Welchen Einfluss hat Query Expansion auf die Performanz von TF-IDF?
7. Inwiefern hängt die Varianz der Kosinus-Ähnlichkeiten mit der Performanz von TF-IDF zusammen?
8. Wie steht diese Arbeit im Vergleich zu anderen IR-Systemen da, die TF-IDF verwenden?
9. Fazit: Was sind die wichtigsten Ergebnisse aus der Evaluation der Anwendung von TF-IDF?

1. Wie gut funktioniert die Anwendung von TF-IDF im gegebenen Kontext?

Im Gesamtdurchschnitt sind die Ergebnisse der Anwendung von TF-IDF[1] in einem Bereich, der deutlich unter den Erwartungen der Zielgruppe bleibt (siehe Abbildung 5) und Tabelle 5. Über die verschiedenen Textebenen zeigt sich ein großer Unterschied in der Performanz der Methode: Die Nutzung der Paragrafen zeigte sich mit einer mittleren Performanz am besten, während insbesondere die Nutzung der Titel schlechte Ergebnisse hervorbrachte. Zusätzlich zeigt sich eine große Variabilität in der Performanz des Ansatzes. Daher repräsentiert der Mittelwert nicht den typischen Wert, sondern das Mittel zweier Extreme: Bei ungefähr einem Drittel der TAQs konnte TF-IDF mindestens für eine der drei Textstellenebenen gute oder sehr gute Ergebnisse erzielen. In den anderen TAQs waren die Ergebnisse häufig sehr gering oder null, dies spiegelt sich auch in den Medianen, die sehr gering ausfallen.

Zwischenfazit TF-IDF erzielte insgesamt unterdurchschnittliche Ergebnisse mit hoher Variabilität: Während bei einigen TAQs gute Resultate möglich waren, blieb die Performanz niedrig, insbesondere bei der Verwendung der Titel.

2. Inwiefern beeinflusst Textlänge die Performanz von TF-IDF beim Ranking?

Um diese Frage zu beantworten, sollen folgende Vergleiche betrachtet werden:

1. Vergleich der Textstellenebenen
2. Vergleich der Textlänge über alle Paragrafen
3. Vergleich der Textlänge über die relevanten Paragrafen und unterschiedlicher Sprachgebrauch

[1] Unter Verwendung der deutschen Studienordnung

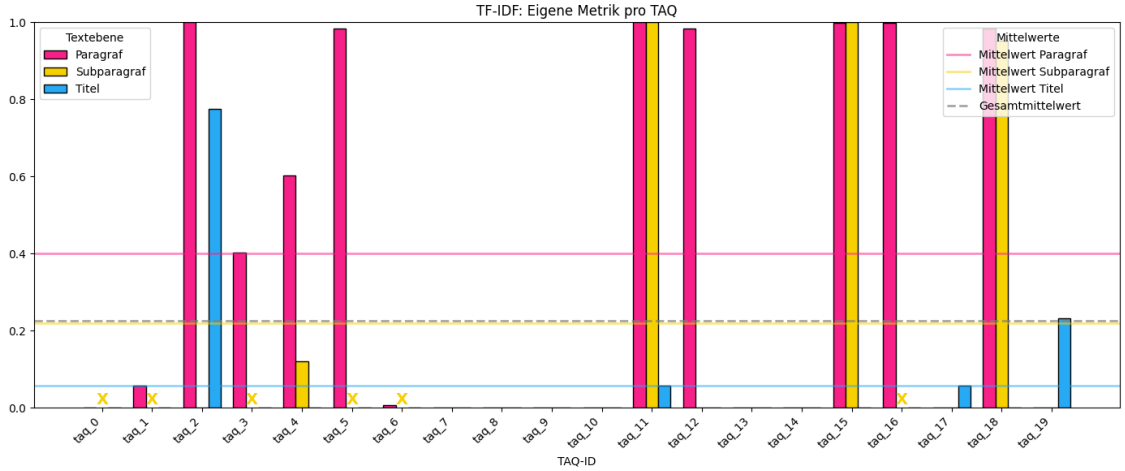

Abbildung 5: *Werte der eigenen Metrik für jede TAQ basierend auf TF-IDF unter Verwendung der originalen Studienordnung.*

1. Vergleich der Textstellenebenen Um die Ergebnisse der drei Textebenen hinsichtlich der Performanz des Ansatzes zu vergleichen, wurde ein Kruskal-Wallis-Test durchgeführt. Die Ergebnisse des Kruskal-Wallis-Tests ergaben eine Teststatistik von $H(2) = 4{,}99$ mit einem p-Wert von $p = 0{,}08$. Dies bedeutet, dass die gefundenen Ergebnisse unter der Annahme, dass es keinen Unterschied macht, welche der Textstellenebene verwendet werden, so auftreten mit 8% gering, jedoch nicht signifikant sind. Die Ergebnisse des Dunn-Post-hoc-Tests mit Bonferroni-Korrektur, der über die Ergebnisse in paarweisen Vergleichen auf signifikante Unterschiede testet, können in Tabelle 6 gefunden werden. Die Wahrscheinlichkeit, dass eine so beobachtete Verteilung zufällig entsteht, unabhängig davon, ob die Paragrafen- oder Titelebene verwendet wird, liegt bei unter 10% und ist damit eher gering, jedoch ebenfalls nicht signifikant. Die schwache Performanz der Titel ist erwartbar, da TF-IDF, wie bereits beschrieben, auf der Übereinstimmung von Wörtern basiert. Da die Titel aus nur wenigen Wörtern bestehen und TF-IDF keine semantische Bedeutung erfassen kann, enthalten die entstehenden Vektoren nur sehr wenige Informationen. Dies verdeutlicht eine zentrale Schwäche des TF-IDF-Ansatzes: In sehr kurzen Texten sind relevante Merkmale oft gar nicht vorhanden oder werden nicht erkannt, da TF-IDF den semantischen Kontext nicht erfassen kann. Trotz des beobachteten Mittelwertunterschieds zwischen Paragrafen und Subparagrafen liefert der p-Wert, da er relativ hoch ist, keinen Hinweis auf einen signifikanten Unterschied zwischen diesen Textstellenebenen. Dies könnte eventuell auf die geringe Stichprobengröße in der Subparagrafenebene ($n = 14$) im Vergleich zur Titelebene ($n = 20$) zurückzuführen sein, was zu einer zu geringen Teststärke geführt haben könnte Kang [2021].

2. Vergleich der Textlänge über alle Paragrafen Um zu untersuchen, wie sich die Performanz von TF-IDF anhand der Textlänge der Paragrafen unterscheidet, sollte geprüft werden, ob durch die Textlänge der durchschnittliche Rang des Textes vorausgesagt werden kann. Da die meisten Paragrafen keine für die TAQs relevante Information enthielten, wurde angenommen, dass sich die Kosinus-Ähnlichkeit für die TAQs gleichmäßig über kurze und lange Textstellen verteilt. Falls besonders häufig Textstellen mit einer bestimm-

Tabelle 5: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch TF-IDF berechneten Ergebnisse für die eigene Metrik*

	M	**SD**	**MD**
gesamt	0,23	0,39	0,00
Textebene			
Paragrafen	0,40	0,47	0,03
Subparagrafen	0,22	0,42	0,00
Titles	0,06	0,18	0,00
Fragenart			
Erweiterte Frage	0,22	0,39	0,00
Faktfrage	0,23	0,42	0,00
Listenfrage	0,23	0,38	0,00

Tabelle 6: *Ergebnisse des Dunn-Post-hoc-Tests mit Bonferroni-Korrektur (p-Werte)*

	Paragrafen	**Subparagrafen**	**Titel**
Paragrafen	1,000	0,415	0,092
Subparagrafen		1,000	1,000
Titel			1,000

ten Länge eine hohe Kosinus-Ähnlichkeit erhalten sollten und somit einen höheren Rang in der Relevanz für eine TAQ, dann könnte das für einen Effekt der Länge der Textstelle sprechen.

Es wurde vermutet, dass sowohl sehr kurze Texte als auch sehr lange Texte weniger häufig hohe Ränge erhalten als mittellange Texte. Aus dem Abschnitt 5.2.1 geht bereits hervor, dass dies für sehr kurze Texte erwartet werden kann. Die zugrunde liegende Annahme für lange Texte ist, dass diese tendenziell eine geringere Informationsdichte aufweisen als kürzere Texte, da mit zunehmender Länge häufiger redundante oder weniger relevante Informationen enthalten sein könnten. Gleichzeitig steigt jedoch auch die Chance, in langen Texten Begriffe zu enthalten, die in einer der TAQs ebenfalls vorkommen. Es wurde vermutet, dass diese Begriffe jedoch insgesamt häufiger vorkommen und daher einen eher geringen TF-IDF-Wert erhalten.

Um das zu prüfen, wurden die Ränge nach Textlängen derjenigen Textstellen nach dem Pre-Processing, die eine Kosinus-Ähnlichkeit ungleich Null aufwiesen ([2] betrachtet. Dies führte zu folgenden Ergebnissen:

Betrachtet man das Streudiagramm in Abbildung 6, so lässt sich für Textlängen bis c.a. 350 Wörtern, auf Paragrafenebene, ein linearer positiver Zusammenhang von Textlänge und mittlerer Rang, an dem der Paragraf gefunden wurde, vermuten. Dies wurde durch eine lineare Regression geprüft, wobei die Werte über einer Textlänge von 350 ausgeschlossen wurden.

[2]In diesem Fall wird der Rang zufällig bestimmt, da es für jede TAQ mehrere Textstellen mit einer Kosinus-Ähnlichkeit gleich Null gibt.

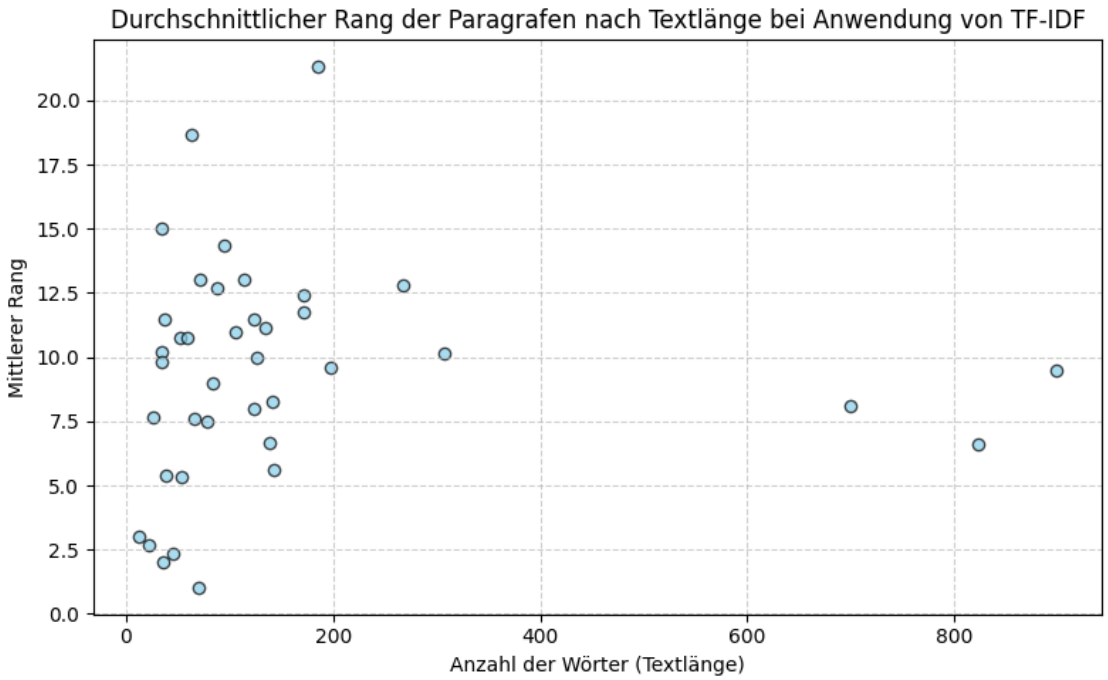

Abbildung 6: *Durchschnittlicher Rang der Paragrafen nach Textlänge bei Anwendung von TF-IDF auf der originales Studienordnung, nach Ausschluss von Kosinus-Ähnlichkeiten mit dem Wert null (Anzahl der Wörter nach dem Pre-Processing).*

Die lineare Regression ergab, dass die Länge des Paragrafen ein signifikanter Prädiktor für den durchschnittlichen Rang der TF-IDF-basierten Retrieval-Ergebnisse war, $F(1, 34) = 4{,}40$, $p = 0,043$, $R^2 = 0,115$, adj. $R^2 = 0,089$. Somit erklärt die Textlänge fast 10% der Varianz der durchschnittlichen Ränge der Paragrafen.

Ein ähnliches Bild ergab sich bei Betrachtung der Subparagrafen. Hier ergab die lineare Regression, bei Ausschluss von Textlängen über 250 Wörtern, dass die Wortanzahl ein signifikanter Prädiktor für den mittleren Rang der abgerufenen Dokumente war, $R^2 = 0,10$, $F(1, 99) = 11,46$, $p = 0,001$. Ein höherer Wortanzahl-Wert sagte einen höheren Rang vorher ($\beta = 0,093$, $SE = 0,028$, $t = 3,39$, $p = 0,001$). Im Anhang ist eine entsprechende Abbildung zu finden (Abbildung 15).

Jedoch passt zu der Vorannahme, dass durch TF-IDF Textstellen mittlerer Länge im Vergleich zu kurzen und langen Texten häufiger höhere Ränge erhalten würden, eine quadratische Regression besser als eine lineare Regression. Daher wurden diesmal die Textstellen mit extrem vielen Wörtern nicht ausgeschlossen. Die quadratische Regression ergab folgende Ergebnisse: Das Modell erklärte etwa 11 % der Varianz des Ranges des Paragrafen ($R^2 = 0,111$, adj. $R^2 = 0,062$).

3. Vergleich der Textlänge über die relevanten Paragrafen und unterschiedlicher Sprachgebrauch Betrachtet man nur die Paragrafen, die relevante Informationen für eine oder mehrere TAQs enthielten, so fällt auf, dass dieselben Paragrafen teils gefunden wurden und teils nicht gefunden wurden (siehe Abbildung 7). Im Folgenden

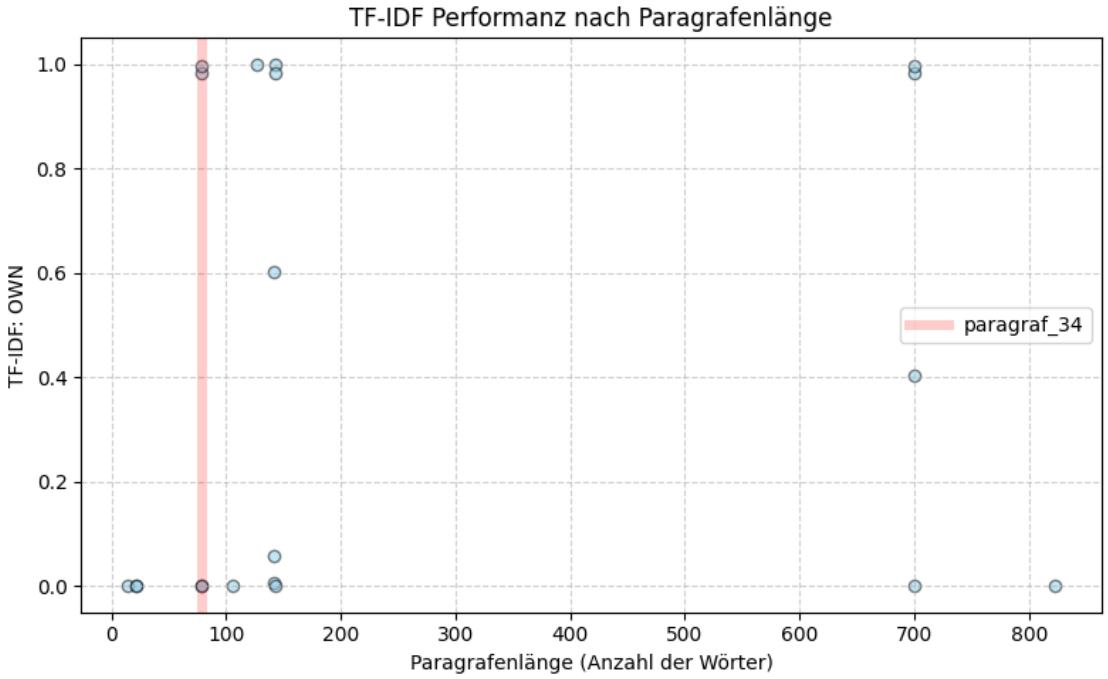

Abbildung 7: *Werte der eigenen Metrik nach Anwendung von TF-IDF auf den Paragrafen, die für eine TAQ die relevante Information enthalten nach Paragrafenlänge (Anzahl der Wörter nach Pre-Processing).*

soll exemplarisch paragraf_34 eingehender betrachtet werden, um mögliche Gründe für dieses Ergebnis zu finden (dieser ist in Abbildung 7 rot markiert). Dieser Paragraf wurde ausgewählt, da er die Antwort für vier TAQs enthält und zweimal gefunden und zweimal nicht gefunden wurde.

Der Text nach dem Preprocessing von paragraf_34 lautet:
(1) masterarbeit masterarbeit nachweis erbracht prufungskandidatin bzw
(2) prufungskandidat lag gestellt thema selbststand wissenschaft method
(3) bearbeit thema masterarbeit fachergrupp gemass anhang entnehm antrag
(4) prufungskandidatin bzw prufungskandidat prufungsausschuss thema
(5) fach zugelass fall prufungskandidatin bzw prufungskandidat glaubhaft
(6) zuweis gestellt thema inhalt bezug nutzung informat genannt
(7) anwendungsgebiet aufweist modul masterarbeit beinhaltet
(8) kolloquium hauptergebnis arbeit verteidigt kolloquium findet
(9) wahl bzw studier entwed bewert masterarbeit statt not
(10) masterarbeit setzt bewert schriftlich arbeit bewert kolloquium
(11) zusamm zulass modul masterarbeit setzt voraus modul
(12) umfang mindest ectspunkte erfolgreich absolviert

Die TAQs, für die dieser Text erfolgreich gefunden wurde, lauten (nach Preprocessing):
taq_12: 'inhalt erwart abschlussarbeit'

taq_15: 'gewicht bewert abschlussarbeit'

Die TAQs, für die dieser Text nicht gefunden wurde, lauten (nach Preprocessing):

taq_13: 'ect erreicht abschlussarbeit anmeld'

taq_14: 'lehrstuhl abschlussarbeit schreib'

In rötlichen Farben sind die Begriffe markiert, die aufgrund eines Vocabulary Mismatch nicht dazu geführt haben, dass die Textstelle als relevant erkannt wurde. In grünlichen Farben entsprechend die Begriffe, die zu einer Übereinstimmung geführt haben. Wie zu erwarten war, war dies nur für die TAQs 12 und 15 der Fall: Der Begriff „Abschlussarbeit" ist ein Überbegriff von sowohl „Masterarbeit" als auch „Bachelorarbeit" und somit kommt es nicht zu einer Übereinstimmung der verwendeten Begriffe. Ähnlich verhält es sich mit dem Begriff „ECTS", der in der TAQ verwendet wird, während in der Studienordnung der spezifischere Begriff „ECTS-Punkte" verwendet wird. Der unterschiedliche Sprachgebrauch zwischen den Studienordnungen und den TAQs könnte also Ursache für das Entstehen des Vocabulary Mismatch sein.

Zwischenfazit

1. Ein statistisch signifikanter Unterschied in der Performanz von TF-IDF zwischen den drei Textebenen konnte nicht gefunden werden. Dennoch deuten die Ergebnisse auf eine tendenziell bessere Performance bei der Nutzung von Paragrafen hin, insbesondere im Vergleich zu den Titeln. Dies lag vermutlich an ihrer Kürze und daran, dass TF-IDF den semantischen Kontext nicht erfassen kann.

2. Vermutlich wird 10% der Varianz des Ranges, an dem die Paragrafen (und Subparagrafen) einsortiert werden durch ihre Textlänge aufklären. Dabei besteht ein quadratischer Zusammenhang, bei dem mittellange Texte häufiger höhere Kosinus-Ähnlichkeiten erhalten.

3. Der unterschiedliche Sprachgebrauch zwischen TAQs und Studienordnung führt zu Vocabulary Mismatches. Begriffe wie „Abschlussarbeit" in den TAQs werden in der Studienordnung durch spezifischere Begriffe wie „Masterarbeit" ersetzt, wodurch relevante Paragrafen von TF-IDF nicht erkannt werden. Diese Schwierigkeiten treten unabhängig von der Textlänge auf, und erklären die Performanz von TF-IDF in einem höheren Ausmaß.

Der Rang einer Textstelle lässt sich nur zu einem geringen Maß durch seine Länge erklären. Auch bei gleicher Länge ist letztendlich die Passung bestimmter Terme der entscheidendere Faktor. Es fehlt jedoch eine umfassendere Analyse in den relevanten Textstellen, da hier nicht genug Daten vorhanden waren.

3. Inwiefern beeinflusst die Art der Frage die Performanz von TF-IDF beim Ranking?

Es zeigten sich über die Fragenarten hinweg sehr ähnliche Mittelwerte (siehe Tabelle 5). Ein dennoch durchgeführter Kruskal-Wallis-Test zeigte, wie erwartet, keinen signifikanten Unterschied in der Performanz zwischen den drei Textebenen, $H(2) = 0{,}44$, $p = 0,80$.

Zwischenfazit Die Art der Frage scheint keinen Einfluss auf die Performanz von TF-IDF zu haben.

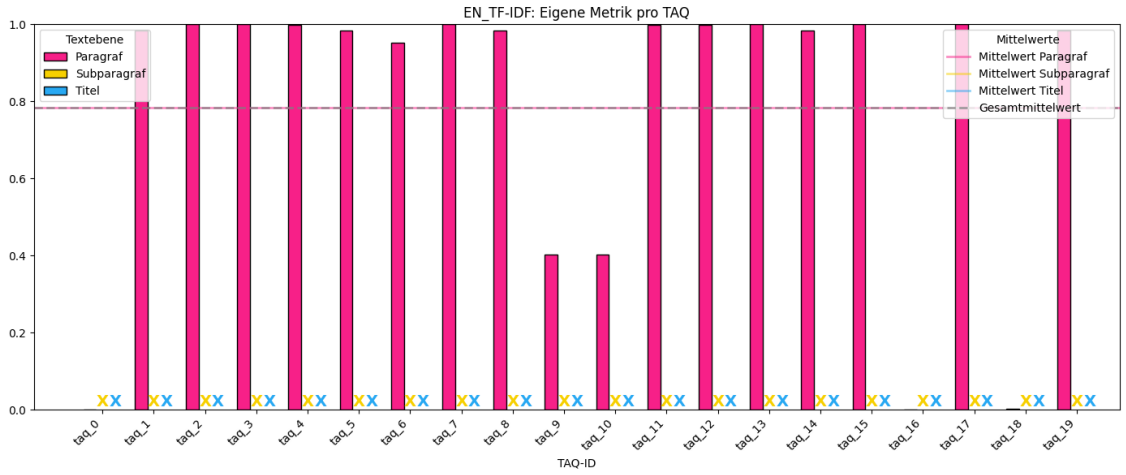

Abbildung 8: *Werte der eigenen Metrik für jede TAQ basierend auf TF-IDF unter Verwendung der ins Englische übersetzten Studienordnung.*

4. Inwiefern verbessert die Übersetzung der Texte ins Englische die Performanz und warum?

Es wurde geprüft, inwiefern die Übersetzung der Studienordnung und der TAQs ins Englische die Performanz von TF-IDF verbessern kann. Dabei wurde angenommen, dass sich durch das Übersetzen sowohl der TAQs als auch der Studienordnung der Einfluss des unterschiedlichen Sprachgebrauches verringert. Im Anhang ist eine beispielhafte Übersetzung einer relevanten Textstelle zu finden (siehe Tabelle 24).

Wie in Abbildung Abbildung 8 zu erkennen, hat sich die Performanz [3] auf der ins Englische übersetzten Studienordnung mit $M = 0{,}78$ ($SD = 0{,}38$) gegenüber der Anwendung auf der ursprünglichen Studienordnung stark verbessert.

Anhand von zwei TAQs soll nun genauer betrachtet werden, wie es zu dieser Verbesserung kam. Ausgewählt wurden die TAQ 13 (*Was sind die inhaltlichen Erwartungen an eine Abschlussarbeit?*) und die TAQ 14 (*Wie viele ECTS müssen erreicht werden, um die Abschlussarbeit anmelden zu können?*). Übersetzt man diese ins Englische, so werden die Begriffe „Abschlussarbeit“ und „ECTS“ zu „thesis“ und „ECTS credits“. Dies hätte in diesem Beispiel das Problem des „Vocabulary Mismatch“ gelöst.

Betrachet man die Performanz über die TAQs weiter, fällt auf, dass einzig für die TAQs taq_16 (*Wie lange ist die Bearbeitungszeit der Abschlussarbeit?* übersetzt zu *How long is the thesis writing period?*) und taq_18 (*Kann man Prüfungen wiederholen, um seine Note aufzubessern?* übersetzt zu *Can I retake exams to improve my grade?*) eine Verschlechterung durch die Übersetzung ins Englische zeigte. Hier könnte durch die Übersetzung der spezifische Gebrauch eines bestimmten Wortes, das zur relevanten Textstelle führt, verloren gehen. Das Wort „Bearbeitungszeit“ wird zu den viel häufiger auftretenden Wörtern „writing“ und „period“, anstelle der Übersetzung „processing time“ (siehe Tabelle Tabelle 7.

[3] Einzig die Paragrafenebene wurde hier betrachtet.

Tabelle 7: *Häufigkeit der Schlüsselbegriffe, die Informationen über die relevante Textstelle enthalten, in Deutsch und Englisch.*

Begriff	**Häufigkeit**
Bearbeitungszeit	9
processing time	3
processing period	3
processing	9
time	21
period	28
writing	10
writing period	0

Zwischenfazit Die Übersetzung ins Englische verbessert insgesamt die TF-IDF-Performanz, da sie Vocabulary Mismatches wie beispielsweise „Abschlussarbeit“ vs. „Masterarbeit“ reduziert. In Einzelfällen kann jedoch durch die Übersetzung der spezifische Gebrauch von Wörtern in TAQ und Studienordnung verloren gehen, was die Performance verschlechtert.

5. Welchen Einfluss hat Stopword-Removal auf die Performanz von TF-IDF?

Auf der englischsprachigen Studienordnung zeigte sich fast kein Unterschied im Mittelwert zwischen der Anwendung mit und ohne Stopword-Removal. Ein Mann–Whitney-U-Test ergab entsprechend einen nicht signifikanten kleineren Wert in der Performanz auf der englischen Studienordnung ohne Stopword-Removal als in der Performanz auf der englischen Studienordnung mit Stopword-Removal, $M_1 = 0{,}77, SD_1 = 0{,}38$, $M_2 = 0{,}78, SD_2 = 0{,}38$, $U = 212{,}0$, $p = 0{,}64$. [4]

Zwischenfazit Im vorliegenden Anwendungskontext scheint Stopword-Removal keinen Einfluss auf die Performanz von TF-IDF zu haben.

6. Welchen Einfluss hat Query Expansion auf die Performanz von TF-IDF?

Wie im Konzept beschrieben (Abschnitt 4.4) wurde Query Expansion auf der ins Englische übersetzten Studienordnung angewendet. Dieses Vorgehen führte zu einer Verbesserung der Performanz. Der Mittelwert über alle TAQs auf der Paragrafenebene verbesserte sich im Vergleich zur Anwendung der in Englisch übersetzten TAQs und Studienordnung von $M = 0{,}78, SD = 0{,}38$ auf $M = 0{,}84$ ($SD = 0{,}36$).

In der Abbildung 8 ist zu erkennen, dass die Methode in dieser Art der Anwendung für fast alle TAQs die relevanten Paragrafen in einem Rang findet, der den Erwartungen an ein gutes System der Nutzenden entspricht. Einzig für die TAQs taq_0 (*Which subjects can I choose?*), taq_16 (*How long is the thesis writing period?*) und taq_18 (*Can I retake exams to improve my grade?*) blieb das Resultat weit unter den Erwartungen der Zielgruppe. Diese drei TAQs erzielten auch bei Verwendung der englischen Versionen, jedoch ohne

[4] Hier wurde lediglich die ins Englische übersetzte Studienordnung verwendet, da es Indizien dafür gab, dass die deutsche Stop Word Liste weniger gut funktioniert hat.

Query Expansion, sehr geringe Werte. Die Verbesserung durch die Query Expansion zeigt sich somit nur für TAQs, die bereits in einem mittleren Bereich funktionierten (betroffen sind taq_1, taq_9 und taq_10).

Ein Mann–Whitney-U-Test ergab keinen signifikant kleineren Wert in der Performanz auf der englischen Studienordnung ohne Query Expansion im Vergleich zur Performanz auf der englischen Studienordnung mit Query Expansion ($M_1 = 0{,}78, SD_1 = 0{,}38$, $M_2 = 0{,}84, SD_2 = 0{,}36$, $U = 180,0$, $p = 0{,}294$).

Zwischenfazit Query Expansion verbesserte die durchschnittliche Performanz von TF-IDF leicht, jedoch nur für TAQs, die bereits mindestens mittelmäßig funktionieren.

7. Inwiefern hängt die Varianz der Kosinus-Ähnlichkeiten mit der Performanz von TF-IDF zusammen?

Die Varianz der Kosinus-Ähnlichkeiten wurde zwischen einer Anwendung von TF-IDF, die mit einem $M = 0{,}23$ (TF-IDF auf der originalen Studienordnung) schlecht funktioniert hat und einer mit $M = 0{,}78$ (TF-IDF auf der englischen Studienordnung) gut funktionierenden verglichen: Hier ergab der t-Test nach Welch einen signifikanten Unterschied der Mittelwerte der Varianzen der Kosinus-Ähnlichkeiten, wobei die Anwendung TF-IDFs mit deutschen TAQs und deutscher Studienordnung kleiner als die Anwendung mit den ins Englische übersetzten Texten war ($t(38) = -2{,}77$, $p = 0{,}0044$). Der Mittelwert betrug in der ersten Gruppe $M = 0,0020$, in der zweiten Gruppe $M = 0,0044$.

Eine lineare Regressionsanalyse wurde durchgeführt, um den Einfluss der Varianz der Kosinus-Ähnlichkeiten auf die Performanz von TF-IDF[5] auf die einzelnen TAQs zu untersuchen. Das Modell war nicht signifikant, $F(1, 14) = 1{,}56$, $p = 0{,}233$, und erklärte nur einen geringen Anteil der Varianz der abhängigen Variable ($R^2 = 0{,}10$, bereinigtes $R^2 = 0{,}036$).

Zwischenfazit Die Varianz der Kosinus-Ähnlichkeiten zeigt einen signifikanten Unterschied zwischen der Anwendung von TF-IDF in einem erfolgreichen Setting und einem nicht erfolgreichen Setting. Jedoch konnte die Kosinus-Ähnlichkeit die Performanz von TF-IDF für die einzelnen TAQs nicht vorhersagen.

8. Wie steht diese Arbeit im Vergleich zu anderen IR-Systemen da, die TF-IDF verwenden?

Für den Vergleich mit anderen Systemen wird die Anwendung von TF-IDF auf der ins Englische übersetzten Studienordnung mit Query Expansion herangezogen, die eigene Metrik ergab einen Wert von $M = ,78$ ($SD = 0{,}38$) [6], der einer guten Performanz mit einem durchschnittlichen Rang von 4 bis 5, an dem die relevante Textstelle gefunden wurde, entspricht. Eine ähnliche Anwendung von TF-IDF in Kombination mit der Kosinus-Ähnlichkeit findet sich bei Yunanda et al. [2022]. Sie nutzten TF-IDF und Kosinus-Ähnlichkeit für ein Empfehlungssystem für Nachrichten und erzielten eine Trefferquote (Hit-Rate) von 80,77

[5] Auf der deutschen Studienordnung und unter Betrachtung lediglich der Paragrafen.

[6] Es wurde ein durchschnittlicher Reciprocal Rank (RR) von $M = 0{,}57$ bei einer Standardabweichung von $SD = 0{,}38$ sowie ein Median von $MD = 0{,}50$ ermittelt.

%. Dazu berechneten sie anhand der Nachrichtenhistorie von Lesern, ob die empfohlenen Nachrichten tatsächlich vom Leser angeklickt wurden. Da sich beide Metriken an der Nutzendenzufriedenheit orientieren und in einem vergleichbaren Bereich liegen, lässt sich festhalten, dass die in dieser Arbeit erreichte Performanz mit derjenigen der Studie von Yunanda et al. [2022] auf einem ähnlichen Niveau liegt.

Zwischenfazit Im Vergleich mit einer ähnlichen Art der Anwendung von TF-IDF in Kombination mit Kosinus-Ähnlichkeit liegt die Performanz der vorliegenden Arbeit in einem ähnlichen Bereich.

10. Fazit: Was sind die wichtigsten Ergebnisse aus der Evaluation der Anwendung von TF-IDF?

Während die Anwendung von TF-IDF auf der originalen, deutschen Studienordnung eine schwache Performanz zeigte, verbesserte sich die Performanz durch die Übersetzung ins Englische deutlich. Die Textlänge scheint im geringen Maße die Performanz von TF-IDF zu beeinflussen und andere Faktoren wie die Art der Frage zeigten keinen Einfluss auf die Performanz. Query Expansion zeigte vor allem für die TAQs, die bereits auf einem mindestens mittleren Niveau funktionierten, eine Verbesserung. Stopword-Removal zeigte einen sehr geringen Einfluss auf die Performanz und inwiefern dieses Ergebnis zufällig entstanden ist, ist unklar. Die Varianz der Kosinus-Ähnlichkeit könnte ein Indikator für die Performanz von TF-IDF zwischen den verschiedenen Anwendungen sein, jedoch nicht auf Ebene der einzelnen TAQs.

5.2.2 Ergebnisse der Anwendung der BERT-basierten Modelle

1. Wie gut funktioniert die Anwendung von BERT-basierten Modellen im gegebenen Kontext?
2. Inwiefern gibt es Merkmale innerhalb der TAQs, die die Performanz der BERT-basierten Modelle beeinflussen?
3. Inwiefern hängt die Varianz der Kosinus-Ähnlichkeiten mit der Performanz von BERT-basierten Modellen zusammen?
4. Wie wirkt sich der Umgang mit limitierenden Context Windows bei längeren Texten auf die Performanz von BERT-basierten Modellen aus?
5. Welchen Einfluss hat die Modellgröße auf die Performanz von BERT-basierten Modellen?
6. Wie effektiv ist der Einsatz von Query Expansion in Kombination mit BERT-basierten Modellen zur Verbesserung der Performanz?
7. Wie steht diese Arbeit im Vergleich zu anderen IR-Systemen da, die BERT-basierte Modelle verwenden?
8. Fazit: Was sind die wichtigsten Ergebnisse aus der Evaluation der Anwendung von BERT-basierten Modellen?

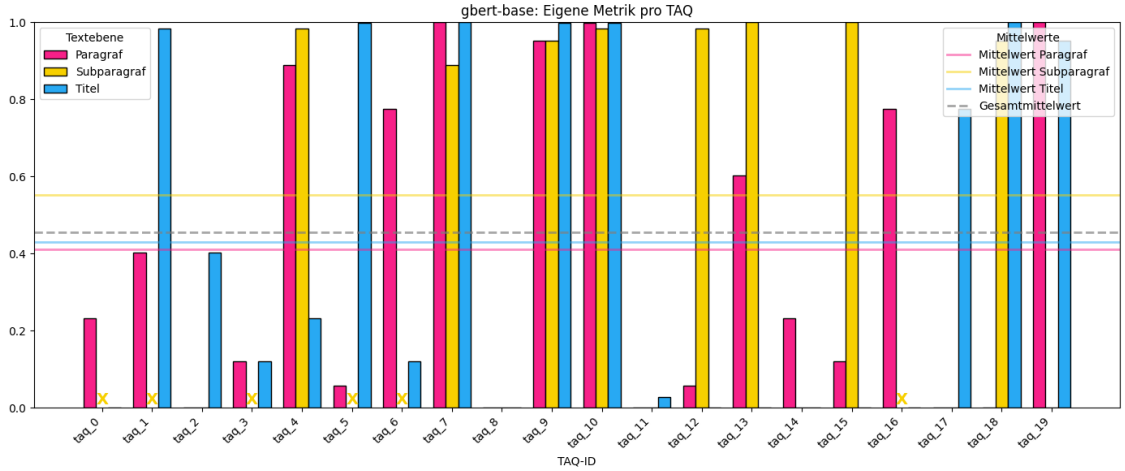

Abbildung 9: *Werte der eigenen Metrik für jede TAQ basierend auf GBERT Base unter Verwendung der originalen Studienordnung.*

1. Wie gut funktioniert die Anwendung von BERT-basierten Modellen im gegebenen Kontext?

Die Ergebnisse der Anwendung der BERT-basierten Modelle können in Tabelle 8 und Abbildung 9 gefunden werden.

GBERT Base Im Gesamtdurchschnitt befinden sich die Ergebnisse der Anwendung des deutschsprachigen Modells gbert-base in einem Bereich, der den mittleren Erwartungen der Zielgruppe entspricht (siehe Tabelle 8 und Abbildung 9). Insgesamt zeigt sich eine große Variabilität in der Performanz des Ansatzes über die verschiedenen TAQs hinweg.

Tabelle 8: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch GBERT berechneten Ergebnisse für die eigene Metrik.*

	Mean	**SD**	**MD**
Gesamt	0,45	0,45	0,23
Textebene			
Paragrafen	0,41	0,41	0,23
Subparagrafen	0,55	0,50	0,92
Titles	0,43	0,46	0,18
Fragenart			
Erweiterte Frage	0,44	0,47	0,23
Faktfrage	0,59	0,44	0,77
Listenfrage	0,27	0,38	0,12

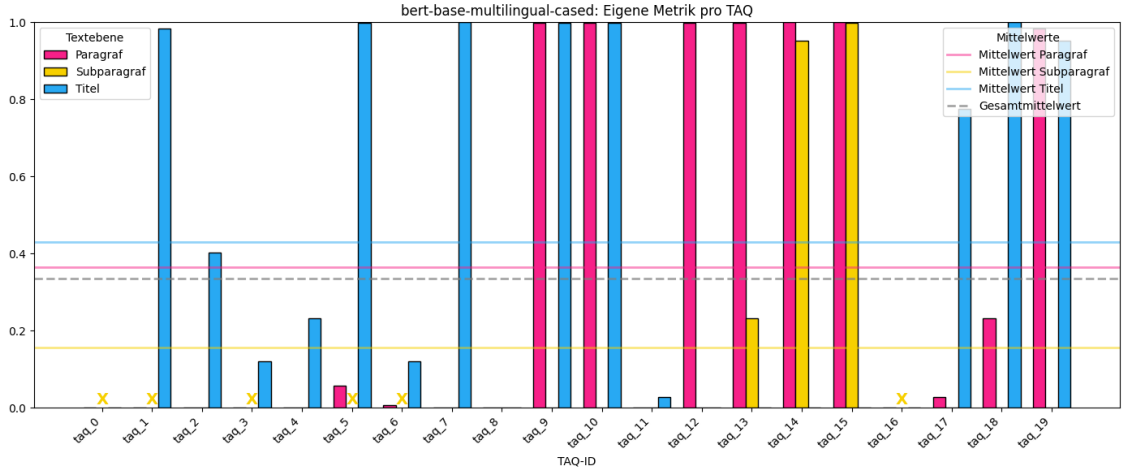

Abbildung 10: *Werte der eigenen Metrik für jede TAQ basierend auf ML BERT Base unter Verwendung der originalen Studienordnung.*

ML BERT Base Die Ergebnisse der Anwendung des mehrsprachigen Modells ML BERT Base befinden sich in einem Bereich, der den unteren Erwartungen der Zielgruppe entspricht (siehe Tabelle 9 und Abbildung 10).

Tabelle 9: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch ML BERT Base berechneten Ergebnisse für die eigene Metrik.*

	M	**SD**	**MD**
Gesamt	0,33	0,45	0,00
Textebene			
Paragrafen	0,36	0,48	0,02
Subparagrafen	0,16	0,35	0,00
Titles	0,43	0,46	0,18
Fragenart			
Erweiterte Frage	0,25	0,43	0,00
Faktfrage	0,38	0,46	0,07
Listenfrage	0,36	0,46	0,06

GELECTRA Base GELECTRA Base wurde lediglich auf Paragrafenebene angewendet, da der Fokus dieser Arbeit auf den Modellen GBERT Base und ML BERT Base lag. Es zeigte sich insgesamt eine geringe Performanz, mit Ergebnissen, die sich in einem Bereich befinden, der den unteren Erwartungen der Zielgruppe entspricht (siehe Tabelle 9 und Abbildung 11).

Vergleich der Modelle GBERT Base zeigte insgesamt eine bessere Leistung im Gesamtdurchschnitt. Besonders deutlich zeigte sich der Unterschied in der Verwendung der

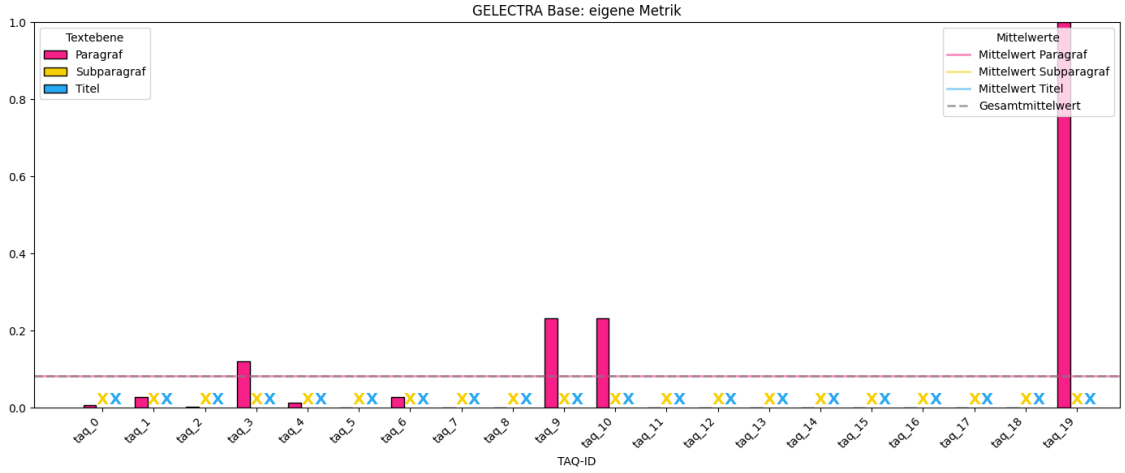

Abbildung 11: *Werte der eigenen Metrik für jede TAQ basierend auf GELECTRA Base unter Verwendung der Paragrafen der originalen Studienordnung.*

Tabelle 10: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch GELECTRA Base berechneten Ergebnisse für die eigene Metrik.*

	M	**SD**	**MD**
Textebene			
Paragrafen	0,08	0,23	0,00
Fragenart			
Erweiterte Frage	0,05	0,09	0,01
Faktfrage	0,03	0,08	0,00
Listenfrage	0,19	0,40	0,00

Subparagrafen. Hier ergab ein Mann–Whitney-U-Test signifikant einen größeren Wert in GBERT Base als in ML BERT Base mit $U = 137{,}0$, $p = 0{,}022$.

Mithilfe von weiteren Mann–Whitney-U-Tests wurde überprüft, ob sich die Mittelwerte zwischen den GBERT Base und den GELECTRA-Modellen signifikant unterscheiden: Der Test ergab einen signifikanten Unterschied zwischen GELECTRA Base und GBERT Base mit ($U = 105{,}5$ und $p = 0{,}01$).

Zwischenfazit Die Ergebnisse zeigen, dass GBERT Base im Vergleich zu den anderen getesteten Modellen die höchste Performanz erzielte und damit den Anforderungen der Zielgruppe am ehesten gerecht wurde. Alle Ergebnisse bleiben jedoch unterhalb der Erwartungen der Stichprobe.

2. Inwiefern gibt es Merkmale innerhalb der TAQs, die die Performanz der BERT-basierten Modelle beeinflussen?

Art der Frage GBERT Base: GBERT Base erzielte bei Faktfragen eine höhere Performanz als bei Listenfragen. Die genauen Ergebnisse sind in Tabelle Tabelle 8 zu finden.

Jedoch zeigte ein Kruskal-Wallis-Test keinen signifikanten Unterschied in der Performanz zwischen den verschiedenen Fragearten, $H(2) = 2{,}08$, $p = 0{,}35$.

ML BERT Base: Bei der Betrachtung der verschiedenen Fragearten zeigte sich die Leistung in den Erweiterten Fragen etwas geringer als in den anderen Fragearten. Die genauen Ergebnisse sind in Tabelle 9 zu finden. Wie sowohl in der Abbildung 10 als auch anhand der SD-Werte aus Tabelle 9 zu erkennen ist, zeigt sich eine große Variabilität in der Performanz des Ansatzes. Um zu überprüfen, ob sich die Werte der eigenen Metrik zwischen verschiedenen Fragenarten signifikant unterscheiden, wurde ein Kruskal-Wallis-Test durchgeführt. Der Test ergab keinen signifikanten Unterschied zwischen den Gruppen, $H(2) = 1{,}77$, $p = 0{,}41$.

Vergleich der Textstellenebenen GBERT Base: Über die verschiedenen Textebenen betrachtet, hat GBERT Base auf der Ebene der Subparagrafen am besten funktioniert und auf Paragrafen- und Titelebene schlechter. Um zu überprüfen, ob sich die Werte der eigenen Metrik zwischen verschiedenen Textebenen signifikant unterscheiden, wurde ein Kruskal-Wallis-Test durchgeführt. Der Test ergab keinen signifikanten Unterschied zwischen den Gruppen, $H(2) = 0{,}03$, $p = 0{,}99$.

ML Bert Base: Über die verschiedenen Textebenen betrachtet, hat ML BERT Base auf der Ebene der Titel am besten funktioniert und besonders schlecht in der Subparagrafenebene. Um zu überprüfen, ob sich die Werte der eigenen Metrik zwischen verschiedenen Textebenen signifikant unterscheiden, wurde ein Kruskal-Wallis-Test durchgeführt. Der Test ergab einen signifikanten Unterschied zwischen den Gruppen, $H(2) = 6{,}99$, $p = 0{,}03$. Es wurde ein Dunn-Test mit Bonferroni-Korrektur für multiple Vergleiche durchgeführt, dabei zeigte sich ein signifikanter Unterschied zwischen der Subparagrafenebene und der Titelebene ($p = 0{,}03$). Die übrigen Paarvergleiche zeigten keine signifikanten Unterschiede ($p > 0{,}05$).

Explorative Analyse der TAQs bei Betrachtung der Paragrafen In Abbildung 12 ist dargestellt, welche TAQs auf Paragrafenebene sowohl für GBERT Base als auch für ML BERT Base jeweils ober- bzw. unterhalb eines Schwellenwerts von 0,5 in der eigenen Metrik lagen. Im Folgenden soll explorativ untersucht werden, welche Gemeinsamkeiten die TAQs aufweisen, die in beiden Modellen entweder oberhalb oder unterhalb dieses Grenzwertes liegen. Ziel ist es, Merkmale zu identifizieren, die potenziell die Performanz der Modelle beeinflussen und vorhersagen könnten.

Bei der Betrachtung der Paragrafen, die für die jeweilige Gruppe an TAQs die relevante Information enthalten, fällt auf, dass die TAQs mit niedrigeren Performanz längere Texte enthalten. Die durchschnittliche Anzahl der Wörter der Paragrafen, die eine niedrigere Performanz zeigten, war $M = 236{,}50$ ($SD = 339{,}41$) und in der mit höheren Performanz $M = 372{,}70$ ($SD = 294{,}47$. Hier handelt es sich jedoch nicht um eine Eigenschaft der TAQ, sondern der zu findenden Textstelle.

Zwischenfazit Es zeigten sich eine leichte, jedoch nicht signifikante, bessere Performanz in den Faktfragen im Vergleich zu den Listenfragen und Erweiterten Fragen. Die Modelle scheinen auf kürzeren Texten besser zu funktionieren; diese Tendenz zeigte sich jedoch nicht bei der Anwendung von ML BERT Base im Vergleich der Textebenen. Es wäre denkbar, dass Faktfragen häufiger durch kürzere Texte beantwortet werden. Die bessere

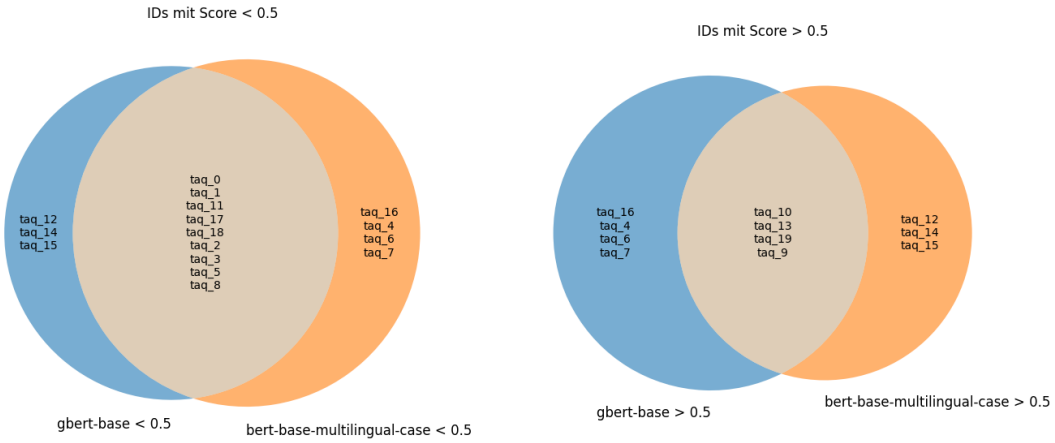

Abbildung 12: *Venn-Diagramme der TAQs auf Paragrafenebene mit Scores unterhalb (links) und oberhalb (rechts) des Schwellenwerts von 0,5 für die Modelle GBERT Base und ML Bert Base.*

Performanz auf Subparagrafenebene im Vergleich zur Paragrafenebene könnte dafür sprechen, dass kürzere Texte mit einer höheren Performanz einhergehen und es dadurch zu einer besseren Performanz der Faktfragen kommt.

3. Inwiefern hängt die Varianz der Kosinus-Ähnlichkeiten mit der Performanz von BERT-basierten Modellen zusammen?

Um zu testen, ob eine größere Varianz in den Kosinus-Ähnlichkeiten mit einer besseren Performanz der Modelle zusammenhängt, wurde das deutschsprachige Modell GBERT Base auf der ins Englische übersetzten Studienordnung angewendet, um eine Verteilung der Cosinus-Werte zu erhalten für eine Anwendung, die eine schlechte Leistung zeigt. Dabei zeigte sich, wie erwartet, eine sehr schlechte Performanz, mit einem $M = 0{,}15$ ($SD = 0{,}31$).[7]

Um zu testen, inwiefern die Standardabweichung der Kosinus-Ähnlichkeit mit der Performanz auf den verschiedenen TAQs zusammenhängt, wurde die Varianz pro TAQ für die beiden Anwendungen GBERT Base auf der deutschen und auf der ins Englische übersetzten Studienordnung berechnet. Anschließend wurde ein gerichteter t-Test [8] nach Welch durchgeführt mit folgendem Ergebnis: Der t-Test nach Welch ergab einen signifikanten Unterschied der Mittelwerte, wobei der Mittelwert der Varianzen aus der Anwendung von GBERT Base auf der englischen Studienordnung kleiner als auf der deutschen Stu-

[7] Überraschenderweise zeigte sich hier ein gutes Ergebnis für die taq_6 (*Can I replace compulsory modules with other modules?*) und taq_7 (*How long can I study before I am de-registered?*). Es könnte sein, dass manche Begriffe im Deutschen und im Englischen relativ ähnlich sind, insbesondere nach dem Pre-Processing, wie zum Beispiel „modules“ mit „Modulen“ und das deutschsprachige Model daher für diese TAQs funktionierte.

[8] Aufgrund angenommener Normalverteilung der Stichproben wurde dieser Test zur Analyse der Mittelwertunterschiede verwendet.

dienordnung war ($t(38)$ = -8,78, p = 0,0000). Der Mittelwert der Varianz betrug in der ersten Gruppe $M = 0,0007$, in der zweiten Gruppe $M = 0,0022$.

Um weiterhin zu prüfen, ob sich die Varianz der Kosinus-Ähnlichkeiten dafür eignet, die Performanz einzelner TAQs vorherzusagen, wurden lineare Regressionen berechnet:

Die Regression der Performanz auf die Varianz der Kosinus-Ähnlichkeiten des GBERT Base Modells in der Anwendung auf die deutsche Studienordnung ergab, dass das Regressionsmodell nicht signifikant war, $F(1, 18) = 0,25$, $p = 0,62$, $R^2 = 0,014$.

Die Regression der Performanz auf die Varianz der Kosinus-Ähnlichkeiten des GBERT Base Modells in der Anwendung auf der ins Englische übersetzten Studienordnung ergab ein marginal signifikantes Regressionsmodell, $F(1, 18) = 3,98$, $p = 0,062$, mit einem Bestimmtheitsmaß von $R^2 = 0,181$. Der Regressionskoeffizient für die Varianz der Kosinus-Ähnlichkeiten betrug $\beta = -713,68$ *(SE* = 357,92), was auf einen negativen Zusammenhang hinweist: Höhere Varianz in den Kosinus-Ähnlichkeiten war tendenziell mit einer geringeren Performanz assoziiert. Dieser Effekt erreichte jedoch nur knapp keine statistische Signifikanz ($p = 0,062$).

Die Regression der Performanz von GBERT Base auf die Varianzen der Kosinus-Ähnlichkeiten zeigte ein scheinbar widersprüchliches Bild: Während der t-Test dafür spricht, dass die englische Version insgesamt eine bessere Performanz zeigt und gleichzeitig weniger Varianz aufweist, zeigte sich in dieser Regression, dass innerhalb der englischen Version die TAQs mit besonders hoher Varianz tendenziell eine schlechtere Performanz zeigen.

Zwischenfazit Eine höhere Varianz in den Cosinus-Werten könnte ein Indikator für bessere Performanz der Modelle sein, allerdings nicht auf TAQ-Ebene, sondern lediglich im Gruppenvergleich verschiedener Anwendungen.

4. Wie wirkt sich der Umgang mit limitierenden Context Windows bei längeren Texten auf die Performanz von BERT-basierten Modellen aus?

Da das Context Window von den auf BERT basierenden Modellen auf 512 Tokens beschränkt ist, muss bei Texten, die länger als dieser Rahmen sind, entschieden werden, wie die Kosinus-Ähnlichkeit berechnet wird. Dazu wird der Text immer zunächst in kleinere Abschnitte (Chunks) aufgeteilt und die Kosinus-Ähnlichkeit für jedes dieser Chunks berechnet. Anschließend wird entweder die größte der Kosinus-Ähnlichkeiten (*Max-Chunks*) für die komplette Textstelle verwendet oder die mittlere Kosinus-Ähnlichkeit (*Mean-Chunks*). Dieses Problem trat nur auf Paragrafenebene auf, da sowohl die Subparagrafen als auch die Titel das Limit von 512 Tokens nicht überschreiten.

Im Vergleich der beiden Möglichkeiten zeigte sich in der Anwendung von GBERT Base mit Max-Chunks bei Betrachtung einzig der Paragrafen ein leicht höherer $M = 0,41$ ($SD = 0,41$) als in der Anwendung mit Mean-Chunks ($M = 0,36$,$SD = 0,42$). Es zeigte sich kein signifikanter Gruppenunterschied (H_0: Bessere Performanz in der Anwendung von Max-Chunks im Vergleich zu Mean-Chunks) bei der Anwendung eines Mann–Whitney-U-Test zwischen den Gruppen, $U = 215,0$, $p = 0,35$.

Zwischenfazit Aufgrund der 512-Token-Grenze von BERT-basierten Modellen wurden lange Textpassagen in kleinere Chunks aufgeteilt und die Kosinus-Ähnlichkeit entweder als

Maximum (*Max-Chunks*) oder Mittelwert (*Mean-Chunks*) berechnet. Der Vergleich beider Methoden mit GBERT Base ergab zwar einen leicht höheren Mittelwert für *Max-Chunks*, jedoch keinen signifikanten Unterschied im Mann–Whitney-U-Test ($p = 0{,}35$).

5. Welchen Einfluss hat die Modellgröße auf die Performanz von BERT-basierten Modellen?

Um zu überprüfen, inwiefern sich der Unterschied zwischen dem Basismodell und einem größeren Modell auf die Performanz auswirkt, wurden die beiden Modelle GBERT und GELECTRA in der Basisversion (base) und größeren Version (large) auf den deutschen Studienordnungen verglichen. Die Kennzahlen der betrachteten Modelle können in Tabelle 11[9] gefunden werden. In Tabelle 12 sind die Mittelwerte und Standardabweichungen der eigenen Metrik für die Paragrafenebene der Modelle zu finden. Ein zweiseitiger Mann–Whitney-U-Test ergab weder einen signifikanten Unterschied zwischen der Leistung von GBERT Large und GBERT Base ($U = 189{,}0$, $p = 0{,}623$), noch zwischen GELECTRA Large und GELECTRA Base ($U = 219{,}0$, $p = 0{,}30$).

Tabelle 11: *Modellarchitektur-Hyperparameter der Base and Large Versionen von GBERT und GELECTRA*

	GBERT		**GELECTRA**	
	Base	**Large**	**Base**	**Large**
Context Window	512	512	512	512
Layers	12	24	12	24
Hidden states	768	1024	768	1024
Attention heads	12	16	12	16
Vocab size (k)	31	31	31	31

Tabelle 12: *Mittelwerte (M) und Standardabweichungen (SD) auf Paragrafenebene für Base und Large Versionen von GBERT und GELECTRA*

Modell	**M**	**(SD)**
$\text{GBERT}_{\text{Base}}$	0,37	(0,41)
$\text{GBERT}_{\text{Large}}$	0,32	(0,41)
$\text{GELECTRA}_{\text{Base}}$	0,08	(0,23)
$\text{GELECTRA}_{\text{Large}}$	0,16	(0,32)

Zwischenfazit Zum Vergleich der Performanz der Basis- und Large-Version wurden die für GBERT und GELECTRA jeweils auf der deutschen Studienordnung getestet, wobei kein signifikanter Unterschied zwischen den Modellgrößen festgestellt wurde. Während sich die Performanz von GELECTRA Base zu Large im Mittelwert verbesserte, zeigte sich in GBERT ein leichter Rückgang der Performanz (vergleiche Tabelle 12).

[9] Tabelle nach Chan et al. [2020]

6. Wie effektiv ist der Einsatz von Query Expansion in Kombination mit BERT-basierten Modellen zur Verbesserung der Performanz?

Dies wurde nur für ML BERT Base[10] getestet: ML BERT Base erzielt mit einem Mittelwert $M = 0{,}66$ und $SD = 0{,}44$ auf Paragrafenebene mit Query Expansion Ergebnisse, die deutlich über denen ohne Query Expansion liegen ($M = 0{,}42$ und $SD = 0{,}47$). Ein Mann–Whitney-U-Test für die Hypothese, dass ML BERT Base mit Query Expansion eine bessere Leistung zeigt, ergibt auf einem 5%-Niveau kein signifikantes Ergebnis ($U = 146{,}5$ und $p = 0{,}074$). Man beachte jedoch den dennoch kleinen p-Wert, der dafür spricht, dass unter Annahme der Gültigkeit, dass es keinen Unterschied zwischen den beiden Gruppen gibt, die beobachtete Verteilung der Werte so auftritt, mit 7% dennoch relativ gering ist.

Ein ähnliches Bild ergab sich im Vergleich der Anwendung von GBERT Base auf der deutschen Studienordnung auf Paragrafenebene: Es zeigte sich eine leichte Verbesserung durch die Verwendung der TAQs samt des Themas ($M = 0{,}41$, $SD = 0{,}41$) im Vergleich zur Verwendung einzig der TAQs ($M = 0{,}37$, $SD = 0{,}41$).

Zwischenfazit Es zeigte sich eine deutliche Verbesserung der Performanz durch Query Expansion, und schon eine leichte Verbesserung der Performanz durch die Verwendung der TAQs samt Themengruppe.

7. Wie steht diese Arbeit im Vergleich zu anderen IR-Systemen da, die BERT-basierte Modelle verwenden?

Um einen Vergleich mit anderen Ansätzen durchzuführen, wurden für GBERT Base der Reciprocal Rank bis Rang 10 berechnet (rr@10). Die Ergebnisse sind in Tabelle 13 zu finden.

Beispielsweise implementierten Nogueira and Cho [2019] BERT Base für query-basierte Passage-Rerankierung und nutzen diese Daten. Anders als in der vorliegenden Arbeit wendeten Nogueira and Cho [2019] Fine-Tuning an: Der genutzte Trainingssatz bestand aus etwa Suchanfragen mit durchschnittlich einer als relevant codierten Textstelle aus dem MS MARCO Datensatz (Bajaj et al. [2016]). Das so für diese Anwendung trainierte BERT Base Modell erzielte einen Mean RR@10 von 34,7. Die in dieser Arbeit erreichte Performanz liegt, insbesondere in der Anwendung auf Subparagrafen und Titeln, in einem vergleichbaren Bereich.

Zwischenfazit Im Vergleich mit einer ähnlichen Art der Anwendung von einem BERT-basierten Modell liegt die Performanz der vorliegenden Arbeit in einem vergleichbaren Bereich.

8. Fazit: Was sind die wichtigsten Ergebnisse aus der Evaluation der Anwendung von BERT-basierten Modellen?

Gbert-base zeigte im Vergleich zu den übrigen Modellen die beste Performanz, jedoch trotzdem niedrige Performanz. Es zeigte sich eine leichte Tendenz hin zu besseren Ergebnissen bei Faktfragen. Die Modelle scheinen insgesamt auf kürzeren Texten besser zu funktionieren, wobei sich dies nicht durchgehend für alle Anwendungen zeigte. Eine höhere

[10] auf Paragrafenebene und unter Verwendung der ins Englische übersetzten Studienordnung

Tabelle 13: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der Ergebnisse der Anwendung von GBERT Base (RR@10) (auf der deutschen Studienordnung)*

	M	**SD**	**MD**
gesamt	0,25	0,34	0,11
Textebene			
Paragrafen	0,16	0,24	0,11
Subparagrafen	0,38	0,43	0,27
Titles	0,24	0,35	0,11
Fragenart			
Erweiterte Frage	0,17	0,28	0,00
Faktfrage	0,26	0,37	0,11
Listenfrage	0,31	0,38	0,12

Varianz in den Kosinus-Ähnlichkeiten könnte ein Hinweis auf eine bessere Modellperformanz sein, jedoch nur im Vergleich zwischen den Modellen, nicht bei Betrachtung einzelner TAQs innerhalb der Anwendung eines Modells. *Max-Chunks* zeigten minimal höhere Werte in der Performanz. Es zeigte sich eine Verbesserung der Performanz durch den Einsatz von Query Expansion. Die Performanz der in dieser Arbeit verwendeten BERT-basierten Modelle befindet sich in einem ähnlichen Bereich zu vergleichbaren Anwendungen.

5.2.3 Ergebnisse der Anwendung von Mistral Large

1. Wie gut funktioniert die Anwendung von Mistral Large im gegebenen Kontext?
2. Inwiefern gibt es Merkmale der TAQs oder der Textstellen, die die Performanz von Mistral Large beeinflussen?
3. Gibt es typische Fehler, die bei der Verwendung von Mistral Large auftreten?

1. Wie gut funktioniert die Anwendung von Mistral Large im gegebenen Kontext?

Mistral Large gab für ca. 43 % der TAQs über die verschiedenen Textebenen die Textstelle zurück, die tatsächlich die Antwort der TAQ enthielt (vgl. Tabelle Tabelle 14). Als richtige Antwort wurde gewertet, wenn das LLM die richtige Textstelle zurückgab, nicht jedoch, wenn direkt die richtige Antwort auf die TAQ zurückgegeben wurde. Dabei wurde im Versuch 1 nur mit den TAQs gearbeitet, dieses Vorgehen wurde zweimal angewendet (Versuch 1.1 und Versuch 1.2). In Versuch 2.1 und Versuch 2.2 wurde zusätzlich zu der TAQ an sich noch die Themengruppe [11] mitgegeben.

Aus der Tabelle 15 für Versuch 1 und Tabelle 16 für Versuch 2 geht weiter hervor, inwiefern die Antworten des LLMs übereinstimmten. Dabei wurde die Übereinstimmung der beiden Durchführungen der Versuche in der Bewertung[12], in der Übereinstimmung der

[11] diese umfassen: „Inhalt des Studiums“, „Dauer des Studiums“, „Abschlussarbeit“ oder „Prüfungen“

[12] Bewertet wurde, ob die relevante Textstelle zurückgegeben wurde.

Tabelle 14: *Anzahl der TAQs mit korrekter Zuordnung durch Mistral Large*

	Versuch 1.1	**Versuch 1.2**	**Versuch 2.1**	**Versuch 2.2**
Paragrafen	8 (40 %)	10 (50 %)	8 (40 %)	10 (50 %)
Subparagrafen	9 (45 %)	8 (40 %)	10 (50 %)	8 (40 %)
Titels	8 (40 %)	8 (40 %)	8 (40 %)	7 (35 %)
Total	25 (41,7 %)	26 (43,3 %)	26 (43,3 %)	25 (41,7 %)

zurückgegebenen Textstelle und in der Übereinstimmung der vermuteten Fehlerursache betrachtet. Man erkennt, dass für die meisten TAQs die Bewertung der zurückgegebenen Textstelle übereinstimmte, lediglich in Versuch 2 auf Paragrafenebene zeigte sich die niedrigste Übereinstimmung mit 60%. Etwas niedriger fallen die Übereinstimmungen in den zurückgegebenen Textstellen aus, mit einer minimalen Übereinstimmung von 45% ebenfalls in Versuch 2 auf Paragrafenebene. Die vermuteten Fehlerursachen stimmen minimal in 56% der TAQs überein (Versuch 2, Subparagrafenebene). Insgesamt zeigt sich eine etwas höhere Übereinstimmung in Versuch 1. In Versuch 1 handelte es sich bei Übereinstimmung in etwa 60% der Fälle[13] um korrekt gefundene Textstellen. Bei Versuch 2 war dies in etwa bei 55% der Fälle* so.

Diese Werte sprechen dafür, dass die zurückgegebenen Textstellen und die Fehlerursachen über-zufällig häufig übereinstimmen und daher als Grundlage für eine Auswertung dienen können. Eine zufällige Übereinstimmung läge bei fünf bis sechs Fehlerkategorien bei etwas weniger als 20%[14].

Tabelle 15: *Übereinstimmungen der Antworten von Mistral Large in Versuch 1 (V1)*

	gleiche Bewertung	*gleiche Textstelle*	*gleiche Fehlerursache**
Paragraf	18 (90 %)	10 (50 %)	6 (60 %)
Subparagraf	18 (90 %)	13 (65 %)	7 (70 %)
Titel	18 (90 %)	16 (80 %)	11 (100 %)

*Nur berücksichtigt, wenn beide Bewertungen falsch waren.

Tabelle 16: *Übereinstimmungen der Antworten von Mistral Large in Versuch 2 (V2).*

	gleiche Bewertung	*gleiche Textstelle*	*gleiche Fehlerursache**
Paragraf	12 (60 %)	9 (45 %)	5 (71 %)
Subparagraf	16 (80 %)	11 (55 %)	5 (56 %)
Titel	17 (85 %)	14 (70 %)	10 (91 %)

*Nur berücksichtigt, wenn beide Bewertungen falsch waren.

[13]Nur Paragrafen und Subparagrafen mit einbezogen.
[14]1 / Anzahl der Kategorien

Zwischenfazit Mistral Large gab für fast die Hälfte der TAQs über die verschiedenen Textebenen die Textstelle zurück, die tatsächlich die Antwort der TAQ enthielt. Es wurden zwei Versuche pro Durchlauf durchgeführt, hier zeigte sich, dass in ebenfalls etwa 50% der zwei Versuche die zurückgegebenen Textstellen übereinstimmten. Es zeigt sich eine minimale Tendenz[15], dass es sich bei Übereinstimmung um die tatsächlich relevante Textstelle handelt. Die vermutete Fehlerursache soll für die weitere Analyse verwendet werden, da sie überzufällig häufig übereinstimmte.

2. Inwiefern gibt es Merkmale der TAQs oder der Textstellen, die die Performanz von Mistral Large beeinflussen?

Es zeigte sich eine leichte Tendenz dahingehend, dass für Listenfragen und Erweiterte Fragen seltener die korrekte Textstelle[16] zurückgegeben wurde. Faktfragen wurden in 6 von 12 der Fälle (50%) mit der korrekten Textstelle beantwortet, Listenfragen in 2 von 7 Fällen (etwa 30%) und Erweiterte Fragen in nur 2 von 8 Fällen (25%). Eine entsprechende **??** findet sich im Anhang.

Betrachtet man die TAQs inhaltlich, so erkennt man folgendes Muster: Die Tags, die die Themen rund um Prüfungen und die Abschlussarbeit umfassen, werden häufiger richtig beantwortet als diejenigen TAQs, die die Themen Inhalt und Dauer des Studiums umfassen. Zugleich sind in den Themen rund um Prüfungen und Abschlussarbeit etwa die Hälfte der TAQs Faktfragen, während in den Themen rund um Inhalt und Dauer des Studiums nur etwa ein Drittel Faktfragen sind.

Sowohl in Versuch 1 als auch in Versuch 2 wurden kürzere Texte häufiger korrekt zurückgegeben (siehe Tabelle 17). Die Texte der nicht korrekt zurückgegebenen Textstellen enthielten im Schnitt etwa doppelt so viele Wörter[17], als die der korrekt zurückgegebenen Texte.

Tabelle 17: *Mittelwert (M) und Standardabweichung (SD) der Textlänge* nach Korrektheit der Antworten von Mistral Large.*

	M	**SD**
Versuch 1		
Korrekt zurückgegeben	186,29	261,95
nicht gefunden	279,60	278,76
Versuch 2		
Korrekt zurückgegeben	111,00	32,00
nicht gefunden	357,86	298,95

*Als Textlänge wurde die Anzahl der Wörter nach Pre-Processing (bei TF-IDF) verwendet und betrachtet wurden nur Paragrafen und Subparagrafen.

[15] Bei Betrachtung der Paragrafen und Subparagrafen
[16] Betrachtet wurden Paragraf oder Subparagraf
[17] nach dem Pre-Processing bei Anwendung von TF-IDF

Tabelle 18: *Fragetypen nach Versuch und Antworttyp (Paragrafen und Subparagrafen) nach Korrektheit der Antworten von Mistral Large.*

	Faktfragen	**Listenfragen**	**Erweiterte Fragen**
Versuch 1			
Korrekt zurückgegeben	4	2	1
Nicht gefunden	3	3	4
Versuch 2			
Korrekt zurückgegeben	2	0	1
Nicht gefunden	3	2	2
Korrekt in %	**50 %**	**29 %**	**25 %**

Zwischenfazit Es scheint eine Tendenz zu geben, dass Faktfragen und Themen rund um Prüfungen und die Abschlussarbeit zusammenfallen und zugleich öfters mit der richtigen Textstelle beantwortet werden, also die Themen rund um Inhalt und Dauer des Studiums, die eher Listenfragen und Erweiterte Fragen umfassen. Außerdem zeigte sich eine Tendenz dahingehend, dass kürzere Textstellen häufiger korrekt zurückgegeben wurden als längere Textstellen. Unklar ist jedoch, ob die Textlänge der beantwortenden Textstelle, die Art der Frage und das Thema unabhängig voneinander die Performanz beeinflussen bzw. wodurch die Performanz an stärksten beeinflusst wird [18].

3. Gibt es typische Fehler, die bei der Verwendung von Mistral Large auftreten?

Im Folgendem werden die angenommenen Fehlerursachen genauer betrachtet, wobei nur bei den TAQs, die in der Wiederholung des Versuches dieselbe angenommene Fehlerursache zeigten, berücksichtigt wurden. Darüber hinaus war in durchschnittlich 17 % der Fälle, in denen ein nicht zur TAQ passender Text zurückgegeben wurde, nicht ersichtlich, warum das LLM zu einem falschen Ergebnis gekommen sein könnte. Diese Fälle wurden auch nicht berücksichtigt. Im Anhang in Tabelle 28 und in Tabelle 29 können die vermuteten Fehlerursachen pro berücksichtigter TAQ für die beiden Versuche gefunden werden.

Folgende Fehlerursachen wurden vermutet:

1. Falscher Kontext
2. Zurückgegebene Textstelle trifft Thema sehr nah, aber enthält die Antwort nicht
3. Titel enthält nicht genug Informationen, um die richtige Textstelle zu finden
4. Antwort ist in der relevanten Textstelle nur implizit enthalten

[18] Die Anzahl der vorhandenen TAQs ist zu gering, um z.B. eine multiple Regression zu berechnen.

Falscher Kontext Ein vermuteter Fehler des LLMs auf Paragrafen- und Subparagrafenebene war die Rückgabe einer Textstelle, die in einem anderen Kontext die TAQ richtig beantwortet hätte. Betrachtet man beispielsweise die taq_4 (siehe Tabelle 5.2.3): Die zurückgegebene Textstelle würde die Frage richtig beantworten, wenn man das „Thema der Masterarbeit“ durch „Fächer außerhalb des Modulhandbuches“ tauschen würde. Mistral Large hat somit eine Textstelle, die eine strukturell gleiche Frage beantwortet, jedoch mit einer Art „Themaverfehlung“ beantwortet.

Diese Art von vermuteter Fehlerursache trat bei vier TAQs in Versuch 1 und bei drei TAQs in Versuch 2 auf. Betrachtet man weiter die tatsächlich relevante Textstelle, dann fällt auf, dass die Antwort auf die TAQ nur implizit enthalten ist (siehe Antwort ist in der relevanten Textstelle nur implizit enthalten).

Tabelle 19: *Gegenüberstellung der von Mistral Large zurückgegebenen Antwort für taq_4 und der tatsächlich relevanten Textstelle.*

taq_4:*
Unter welchen Bedingungen kann ich Fächer außerhalb des Modulhandbuches wählen?

Antwort von Mistral Large: paragraf_35, subparagraf_1	**Relevante Textstelle: paragraf_39, subparagraf_2**
Die Antwort auf die Frage, unter welchen Bedingungen man Fächer außerhalb des Modulhandbuches wählen kann, befindet sich im Text 106: Das Thema der Masterarbeit ist aus einer Fächergruppe gemäß Anhang 2 zu entnehmen. Auf Antrag der Prüfungskandidatin bzw. des Prüfungskandidaten kann vom Prüfungsausschuss auch ein Thema aus einem anderen Fach zugelassen werden. In diesem Fall ist von der Prüfungskandidatin bzw. vom Prüfungskandidaten glaubhaft nachzuweisen, dass das gestellte Thema inhaltlich einen Bezug zur Nutzung der Informatik in den genannten Anwendungsgebieten aufweist.	Die Modulgruppe A2 bietet die Möglichkeit, je nach Interesse bestimmte Studien-schwerpunkte zu setzen. Im Wahlpflichtbereich Angewandte Informatik (AI) stehen abhängig vom aktuellen Lehrangebot Module der Fächer gemäß Anhang 2 a) zur Auswahl. Im Wahlpflichtbereich Anwendungskontext und Überfachliche Qualifikationen können ausgewählte Module aus dem Angebot der Informatik und Wirtschaftsinformatik gewählt werden sowie Module zur fachbezogenen Informationsverarbeitung aus dem Angebot anderer Fakultäten. Weitere Veranstaltungen zu Ethik und Datenschutz runden das Angebot in Modulgruppe A2 ab.

* Subparagrafenebene, Versuch 1

Antwort ist in der relevanten Textstelle nur implizit enthalten Ein vermuteter Fehler des LLMs auf Paragrafen- und Subparagrafenebene war die Rückgabe einer irrelevanten Textstelle, während die relevante Information in einer Textstelle nur implizit enthalten war. Betrachtet man beispielsweise die taq_4 (siehe Tabelle 5.2.3): Die Antwort auf die Frage, inwiefern Fächer außerhalb des Modulhandbuches wählbar sind, wird nur implizit von der Studienordnung beantwortet: Im Bereich A2 ist es möglich, Module aus dem gesamten Angebot der Informatik und Wirtschaftsinformatik zu wählen, und somit

auch solche Module, die nicht zwingend im Modulhandbuch des Studiengangs Computing in the Humanities gelistet sind. Für die Erschließung dieser Information braucht es das Hintergrundwissen, dass im Modulhandbuch nicht zwingend alle Module des gesamten Angebots der Informatik und Wirtschaftsinformatik gefunden werden können und es die Möglichkeit gibt, Module per Antrag zu belegen.

Zurückgegebene Textstelle trifft Thema sehr nah, aber enthält die Antwort nicht Ein häufig vermuteter Fehler[19] ist die Rückgabe einer Textstelle, die thematisch sehr eng mit der TAQ verbunden ist, jedoch nicht die Information enthält, die die TAQ tatsächlich beantwortet hätte.

Zum Beispiel bei der Beantwortung der taq_5 (siehe Tabelle 5.2.3) gibt das Mistral Large eine Textstelle zurück, die die Information gibt, welche thematischen Module in der Modulgruppe A1 zu wählen sind. Mistral antwortet somit zu allgemein: Es wird die Information gegeben, aus welchen Bereichen („Informatik und Angewandte Informatik “) gewählt werden kann, anstelle der Nennung der konkreten Module. Außerdem fällt auf, dass die tatsächlich relevante Textstelle die TAQ umfassend, d.h. für alle möglichen Modulgruppen beantworten kann, während die zurückgegebene Textstelle die (zu ungenaue) Information nur für die Modulgruppe A1 enthält.

Mistral scheint also nicht in der Lage zu sein, zu erkennen, ob

1. die TAQ spezifisch genug beantwortet wurde, und
2. die TAQ umfassend beantwortet wurde.

Tabelle 20: *Gegenüberstellung der von Mistral Large zurückgegebenen Antwort für taq_5 und der tatsächlich relevanten Textstelle.*

taq_5:*
Welche Module sind in den Modulgruppen wählbar?

Antwort von Mistral Large: paragraf_37, subparagraf_0	**Relevante Textstelle: anhang_0**
116: (2) In Modulgruppe A1 werden Grundlagen in Informatik und Angewandter Informatik gelegt, die für die übrigen Modulgruppen benötigt werden. Vorkenntnisse in Informatik sind dabei nicht zwingend erforderlich.	Module und Modulgruppen des Masterstudiengangs Computing in the Humanities ... B. Modulgruppen ... **

* Subparagrafenebene, Versuch 1
** Es folgen verschiedene Tabellen mit den Modulen je nach Modulgruppe, ein Beispiel ist im Anhang zu finden in Abbildung 24

Titel enthält nicht genug Informationen, um die richtige Textstelle zu finden In fast der Hälfte der Fälle bei Verwendung des Titels konnte das Mistral Large die korrekte Textstelle nicht finden. Die vermutete Fehlerursache ist immer, dass der Titel nicht

[19] Diese Fehlerursache wurde in Versuch 1 viermal und in Versuch 2 dreimal vermutet

genügend Informationen enthält, um die TAQ zu beantworten. Somit ist interessanter zu betrachten, unter welchen Bedingungen der Titel zur tatsächlich relevanten Textstelle geführt hat. Hier sollen nur die Fälle betrachtet werden, die in Versuch 1.1 und Versuch 1.2 und genauso in Versuch 2.1 und Versuch 2.2 in der Rückgabe derselben Textstelle übereinstimmten. Im Anhang in Tabelle 30 sind die Titel samt der Information, ob sie richtig oder falsch erkannt wurden, aufgelistet.
Es fällt auf, dass die folgenden zwei Titel sowohl falsch als auch korrekt erkannt wurden:

1. Studiendauer und Studienumfang (paragraf_29)
2. Anerkennung von Studienzeiten, Prüfungsleistungen und Praktikumsleistungen (paragraf_5)

Der Titel „Studiendauer und Studienumfang“ wurde für folgende TAQ korrekt erkannt:

1. *Wie lange kann ich studieren, bis ich exmatrikuliert werde?* (taq_7)

und für folgende TAQs fälschlicherweise zugeordnet:

1. *Wie erhalte ich eine Studienhöchstzeitverlängerung?* (taq_8)

   Korrekter Titel: „Endgültiges Nichtbestehen“ (paragraf_19)

2. *Muss ich die insgesamt 180 ECTS bzw. 120 ECTS genau erreichen?* (taq_9)

   Korrekter Titel: „Wiederholung der Bachelor- oder Masterarbeit“ (paragraf_18)

3. *Was passiert, wenn ich mehr als 180 ECTS bzw. 120 ECTS erreiche?* (taq_10)

   Korrekter Titel: „Wiederholung der Bachelor- oder Masterarbeit“ (paragraf_18)

Hier zeigt sich, dass der Titel „Studiendauer und Studienumfang“ als ein Überbegriff der Titel, die nicht korrekt erkannt wurden, aufgefasst werden könnte. Das könnte für eine Tendenz von Mistral Large sprechen, im Zweifel eine allgemeinere Antwort einer spezifischeren Antwort zu bevorzugen.

Zwischenfazit In etwa 17% der Fälle konnte keine Fehlerursache vermutet werden, ansonsten können typische Fehlerursachen angenommen werden. Besonders häufig wurden Textstellen zurückgegeben, die zwar eng mit dem Thema der TAQ zusammenhingen, jedoch die Antwort nicht enthielten. Außerdem wurde der Kontext verwechselt, und insbesondere bei der Verwendung der Titel zeigte sich, dass diese häufig nicht genügend Informationen enthalten, um auf die Informationen im Paragrafen zu schließen. Die Ergebnisse könnten dafür sprechen, dass Mistral Large Schwierigkeiten damit hat, den richtigen Detailgrad zu treffen, mit einer Tendenz, allgemeinere Texte spezifischeren Texten im Zweifel zu bevorzugen.

5.3 Ergebnisse aus der Nachbereitung

5.3.1 Vergleich der Ansätze

Es zeigte sich, dass TF-IDF unter bestimmten Bedingungen eine sehr gute Performanz zeigte, die den Erwartungen der Zielgruppe am ehesten entsprechen würde. Die Performanz der BERT-basierten Modelle bleibt hinter der Performanz von TF-IDF und Mistral Large: Mistral Large gab für etwa 50% der Fälle die relevante Textstelle zurück, während bei TF-IDF für etwa 75% der TAQs und bei den BERT-basierten Modellen in etwa 40% der TAQs die relevante Textstelle in den oberen Rängen gefunden wurde [20].

Die Ergebnisse deuten darauf hin, dass die Verwendung von TF-IDF zu bevorzugen ist, jedoch nur mit Übersetzung der Texte und Anfragen ins Englische. Dies ist durchaus überraschend, da TF-IDF nicht in der Lage ist, den semantischen Kontext, anders als die BERT-basierten Modelle oder das LLM Mistral Large, zu erfassen.

TF-IDF bietet folgende Vorteile:
- Determiniertheit
- Nachvollziehbarkeit der Ergebnisse
- Erklärbarkeit der Ergebnisse
- Geringer Ressourcenverbrauch und schnelle Berechnung

Ein Nachteil ist jedoch, dass das Übersetzen ins Englische diese Vorteile zum Teil wieder aufwiegt. Jedoch sind die Ergebnisse dennoch besser nachvollziehbar und erklärbar als die Ergebnisse der anderen in dieser Arbeit verwendeten Ansätze.

Herausforderungen des Anwendungskontext

Es gibt spezifische Herausforderungen des Anwendungskontextes, im Folgenden sollen diese erläutert werden und betrachtet werden, inwiefern die Ansätze mit diesen umgehen könnten und welche Implikationen sich daraus ableiten lassen.

1. Die deutsche Sprache
2. Unterschiedlichen Längen der Textstellen und geringer Informationsgehalt der Titel
3. Unzureichende Ausrichtung der Studienordnung auf Studierende
4. Herausforderungen der Ansätze

Die deutsche Sprache Aus Tabelle 4 geht hervor, dass die ins Englische übersetzte Studienordnung mehr Wörter umfasste und zugleich weniger einzigartige Begriffe. Vermutlich liegt dies an den zusammengesetzten Wörtern im Deutschen. Dies stellt insbesondere eine Herausforderung für TF-IDF dar, da zusammengesetzte Wörter im Deutschen relativ spezifisch sind und ihre semantische Nähe zu ähnlichen Begriffen bei Verwendung von TF-IDF nicht abgebildet werden kann. Dies konnte jedoch mit Hilfe der Übersetzung ins Englische abgemildert werden.

[20] Hier wurden immer die besten Ergebnisse der Ansätze betrachtet.

Um sicherzugehen, dass es sich hier um einen Effekt der Sprache und nicht des Übersetzens an sich handelt, wurde der Text der ins Englische übersetzten Studienordnung wieder ins Deutsche übersetzt. Die folgenden Word-Clouds im Anhang (siehe Abbildung 16, Abbildung 17 und Abbildung 18) stellen die Wörter nach Häufigkeit des Auftretens[21] dar. Es zeigte sich, dass in der originalen Studienordnung *bzw* nicht als Stoppwort erkannt wurde.

Für die Anwendung von TF-IDF einzig auf den Paragrafen der ins Deutsche rückübersetzten Studienordnung ergaben sich folgende Werte: $M = 0{,}39$, $SD = 0{,}49$[22]. Bei der Verbesserung der Performanz auf der ins Englische übersetzten Studienordnung scheint es sich somit tatsächlich um einen Effekt der Sprache, nicht des Übersetzens an sich zu handeln.

Implikationen:

1. Sprachspezifische Eigenschaften des Englischen könnten erklären, warum die Übersetzung der Studienordnung die Performanz von TF-IDF verbessert.

Unterschiedlichen Längen der Textstellen Die Textlänge scheint alle Ansätze zu einem mindestens geringen Maß zu beeinflussen. Die Verwendung der Titel zeigte sich immer als unpassend, da in den Titeln nicht genügend Informationen enthalten waren, um auf die richtige Textstelle zu schließen. Diese Herausforderung gilt nicht nur für die betrachteten Ansätze, sondern vermutlich auch für Studierende, die nach einer Information in den Studienordnungen suchen und sich im Suchprozess eventuell an den Titeln orientieren.
Die in dieser Arbeit betrachteten Ansätze zeigten eine Tendenz dahingehend, dass:

- **TF-IDF** benötigt eine Mindestanzahl an Wörtern, um funktionieren zu können. Sobald diese überschritten ist, ist die Performanz nur im geringen Maß von der Textlänge beeinflusst. Einen größeren Einfluss haben andere Faktoren, wie die Passung des Sprachgebrauchs zwischen der Anfrage und den Texten (siehe Abschnitt 5.3.1).
- **BERT-basierte Modelle** scheinen für TAQs, deren relevante Textstelle länger ist, schlechter zu funktionieren. Bei Texten, die über das Context-Window hinausgehen, muss mit Chunks gearbeitet werden.
- **Mistral Large** scheint für TAQs, deren relevante Textstelle länger ist, schlechter zu funktionieren[23]. Bemerkenswert ist, dass Mistral Large bei Verwendung einzig der Titel nicht wesentlich schlechter funktionierte als bei Verwendung der längeren Textebenen.

Implikation:

1. Da es eine Tendenz bei den (L)LMs zu geben scheint, für kürzere Texte bessere Ergebnisse zu erzielen, bietet es sich an, Texte so vorzubereiten, dass sie die nötigen

[21] Größere Wörter traten häufiger auf.

[22] Auf der ins Englische übersetzten Studienordnung: $M = 0{,}78$, $SD = 0{,}38$, und der originalen Studienordnung: $M = 0{,}40$, $SD = 0{,}47$

[23] Es kann jedoch nicht ausgeschlossen werden, dass andere Faktoren wie Art der Frage oder Thema der TAQ einen größeren Einfluss haben. Dasselbe gilt für BERT-basierte Modelle.

Informationen erhalten und zugleich kürzer sind. Da TF-IDF in der Anwendung wesentlich weniger Ressourcen benötigt als die (L)LMs, könnten die Texte mit Hilfe mit diesem Ziel vorverarbeitet werden (siehe Unterabschnitt 5.3.2).

2. Aus der Perspektive der Suchenden stellt sich die Frage, inwiefern Suchende kürzere Textstellen, in denen sie die gesuchte Information schnell finden können, oder längere Textstellen, die möglicherweise wichtige Informationen für den Kontext ihrer Frage enthalten, bevorzugen, könnte in einer weiteren Umfrage herausgefunden werden. Dann könnte man die Ansätze für eine bestimmte Textlänge optimieren.

Unzureichende Ausrichtung der Studienordnung auf Studierende Bei den Studienordnungen handelt es sich um Texte, die einerseits in einem juristischen Format formuliert sind und somit das Ziel haben, einen Rechtsrahmen zu legen und andererseits als Informationsquelle für Studierende zur Verfügung stehen sollen. Studienordnungen befinden sich somit in einem Interessenskonflikt bezüglich der Art der Formulierungen und dem Inhalt der Informationen, denn Rechtstexte richten sich sowohl an Institutionen, die für ihre Umsetzung zuständig sind, als auch an Bürger als juristische Laien, wobei beide Gruppen unterschiedliches Hintergrundwissen mitbringen und die Texte mit anderen Absichten interpretieren (Becker [2001]). Manche Informationen sind dadurch nur implizit in den Studienordnungen zu finden. Interessanterweise scheint dies kein Problem bei der Anwendung von TF-IDF darzustellen, sehr wohl aber bei der Anwendung von Mistral Large. Betrachtet man ausschließlich jene TAQs, bei denen bei der Anwendung von Mistral Large diese spezifische Fehlerursache vermutet wurde, und analysiert die Performance der BERT-basierten Methoden in genau diesen Fällen, ergibt sich ein uneinheitliches Bild: Während gbert-base die korrekte Textstelle häufig in einer hohen Platzierung identifizieren konnte, gelang dies bert-base-multilingual-cased hingegen nicht.

Darüber hinaus wurden TAQs gefunden, deren Antwort nicht in den Studienordnungen zu finden war [24] Daher wird das Informationsbedürfnis bei lediglich der Verwendung der Studienordnungen nicht abgedeckt. Folglich müssten getestet werden, inwiefern die Ansätze auf andere Dokumente, wie die Modulhandbücher, anwendbar sind.
Implikationen

1. Die beschriebenen Herausforderungen könnten durch eine Verbesserung der Formulierung der Studienordnungen verringert werden, und natürlich durch eine stärkere Werbung für die Studienberatung, insbesondere in Englisch: In der Umfrage 1, in der lediglich Studierende der Universität Bamberg teilnahmen, zeigte sich in der Gruppe derjenigen, die an der englischsprachigen Umfrage teilnahmen, eine Tendenz für eine geringere Teilnahme an einer Studienberatung.

2. Anstelle einzig der Verwendung der Studienordnungen sollten die Studienordnungen in Kombination mit den Modulhandbüchern verwendet werden. Studierende scheinen Schwierigkeiten in der genauen Unterscheidung der unterschiedlichen Dokumente zu haben.

[24] Diese wurden nicht verwendet.

Herausforderungen der Passung zwischen den Studienordnungen und TAQs

Es zeigt sich, dass die TAQs in einer Art und Weise formuliert sind, die näher am Sprachgebrauch der Studierenden ist als am juristisch geprägten Sprachgebrauch der Studienordnung. Zugleich sind die Begriffe der TAQs tendenziell allgemeiner formuliert. Durch das Übersetzen in eine andere Sprache, hier Englisch, werden diese Unterschiede verringert.

Herausforderungen der Ansätze

Die allgemeinen Vor- und Nachteile der verschiedenen Ansätze zeigen sich auch in dem vorliegenden Anwendungskontext, so sind die Ergebnisse von TF-IDF wesentlich besser nachvollziehbar als die Ergebnisse von den BERT-basierten Modellen und den Ergebnissen von Mistral Large. Die BERT-basierten Ergebnisse lassen sich beispielsweise durch die Möglichkeit, die Tokenisierung einzelner Wörter nachzuvollziehen, noch besser interpretieren als die Ergebnisse von Mistral Large. Der Rechenaufwand ist für TF-IDF wesentlich geringer als der für die BERT-basierten Modelle, die wesentlich länger in der Anwendung brauchen[25] Bei der Anwendung von LLMs besteht überdies immer die Gefahr der Halluzinationen (Liu et al. [2024]), was im vorliegenden Kontext besonders negativ ins Gewicht fallen könnte, wenn Studierende auf Grund falscher Informationen Entscheidungen treffen würden.

5.3.2 Kombination der Ansätze

Da die Bert-basierten Modelle und Mistral Large gut für kurze Texte funktionierten, könnten unter Verwendung von TF-IDF Begriffe, die für die betrachtete Textstelle eine hohe Spezifität aufweisen, herausgefunden und in Kombination mit GBERT und Mistral Large verwendet werden. Anstelle der Textstelle könnten dann lediglich die Begriffe, die diese gut repräsentieren, verwendet werden.

Dies wurde explorativ für die ins Englische übersetzten Studienordnung unter Verwendung von BERT ML unter Verwendung der Begriffe (aus den Paragrafen), mit einem TF-IDF Wert, der mindestens 0,30 entsprach, getestet. Dieser Ansatz erzielte einen $M = 0{,}23$ ($SD = 0{,}39$) und konnte somit BERT ML auf Paragrafenebene ($M = 0{,}15$, $SD = 0{,}31$) übertreffen.

[25] Für die vorliegende Arbeit wurden die Modelle in Kaggle`https://www.kaggle.com/` implementiert. Die Dauer für die Anwendung von BERT-basierten Modellen variierte stark, lag jedoch bei mindestens 20 Minuten und brach häufiger ab.

Kapitel 6

Diskussion & Fazit

6.1 Diskussion des Empirischen Teils

6.1.1 Umfrage 1

Insgesamt führte die Umfrage 1 zum gewünschten Erfolg, es konnten insgesamt 20 TAQs, die die Studienordnung betrafen, zu unterschiedlichen Themen herausgefunden werden. Dennoch ist unklar, ob die gefundenen TAQs das Informationsbedürfnis der gesamten Studierenden abdecken. Die Gruppe derjenigen, die auf Englisch studieren, die im Bachelor oder am Anfang ihres Studiums stehen, sowie Personen mit Fragen zu Nachteilsausgleichen waren unterrepräsentiert. Obwohl Begriffe wie „Studien- und Fachprüfungsordnung“ erklärt wurden, sprechen die Ergebnisse dafür, dass diese zumindest zum Teil mit z.B. den Modulhandbüchern verwechselt wurden. Dies kann jedoch auch als Ergebnis gewertet werden, da Studierende hier scheinbar Schwierigkeiten haben, zu unterscheiden. Die Einteilung der gefundenen TAQs nach Art der Frage und nach dem Thema wurde nach bestem Wissen durchgeführt, hier wäre es jedoch angemessener gewesen, weitere Personen herbeizuziehen, um eine höhere Reliabilität zu erreichen.

6.1.2 Umfrage 2

Mit insgesamt über 60 Teilnehmenden konnte in Umfrage 2 eine große Stichprobe erzielt werden, die eine gute Basis für die Entwicklung der eigenen Metrik lieferte. Daher, dass es keinen Unterschied zwischen den Personen, die zum Zeitpunkt der Umfrage studierten, und den Personen, die sich zum Zeitpunkt nicht in einem Studium befanden, gab, kann vermutet werden, dass die Ergebnisse relativ gut generalisierbar sind. Jedoch wurden, da dies nicht der Fokus der Arbeit war, weitere Personenmerkmale nicht erhoben und somit lässt sich nicht genau abschätzen, wie heterogen die Stichprobe zusammengesetzt war. Inhaltlich wurde jedoch in dem Fragebogen sichergestellt, dass die Ergebnisse zumindest auf die Anwendung eines Systems, das Modulhandbücher verwendet, angewendet werden können, da explizit auch nach Modulhandbüchern gefragt wurde. In den Freitextanmerkungen wurde u.A. erwähnt, dass es schwierig sei, sich vorzustellen, mit welchem Rang man zufrieden wäre. Diese Anmerkung könnte darauf hindeuten, dass die Umfrage nicht eindeutig genug formuliert gewesen ist und ein Beispiel hilfreich gewesen wäre. Ursprünglich wurde sich explizit gegen ein Beispiel entschieden, um die Antworten der Befragten nicht zu beeinflussen.

Betrachtet man die Ergebnisse aus Umfrage 2, so passen diese zu den Ergebnissen von Joachims et al. [2017]: In Umfrage 2 lag der Median nach der Frage, an welcher Stelle das relevante Dokument maximal stehen darf, damit die Person mit dem IR-System noch zufrieden ist, zwischen 3 und 4. Damit übereinstimmend fanden Joachims et al. [2017] durch Eye-tracking Experimenten heraus, dass ungefähr die Hälfte aller Suchenden nur die ersten drei vorgeschlagenen Dokumente prüften.

6.2 Diskussion der Evaluationsmethoden

6.2.1 Diskussion der eigenen Metrik

Eine Metrik zur Evaluation eines IR-Systems sollte sensibel gegenüber möglichen Qualitätsunterschieden verschiedener Systeme sein und diese korrekt abbilden können (Mandl [2010]). Hier ist relevant, ob die Qualität eines Systems aus einer nutzenden-zentrierten oder einer rein systeminternen Perspektive bewertet wird (v.g.l. Wicaksono and Moffat [2020]). Die anhand der Umfrage 2 entwickelte Metrik hat zum Ziel, das IR-System unter Verwendung der Zufriedenheit der Nutzenden als Qualitätsmerkmal zu bewerten. Dieses Ziel ist durch eine Verankerung in der Zielgruppe gelungen. Betrachtet man Abbildung 2 so wird deutlich, dass die Metrik lediglich für Ränge bis etwa 12 zwischen den Rängen unterscheidet und bildet dies entsprechend ab. Das ist aus einer nutzenden-zentrierten Perspektive so gewollt, da in Umfrage 2 deutlich wurde, dass, sobald eine relevante Textstelle in einem hohen Rang gefunden wurde, das System als gescheitert gelten kann. Zugleich stellt es bei der Anwendung von TF-IDF aus system-zentrierter Perspektive kein Problem dar, da für die höheren Ränge die Kosinus-Ähnlichkeiten null sind und die Ränge dann zufällig zugewiesen werden. Eine Metrik, die über diesen Rang hinaus die unterschiedlichen Ränge bewerten würde, könnte daher zu falschen Schlüssen führen. Nachteile der Verwendung einer zielgruppenorientierten Metrik sind die Schwierigkeit, die erreichten Werte mit Werten anderer Systeme, die andere Metriken verwenden, zu vergleichen. Sobald die Zufriedenheit der Nutzenden in anderen Systemen explizit berichtet wurde, gelingt dies jedoch wieder.

6.2.2 Statistische Tests

Inferenz-statistische Tests wurden in dieser Arbeit verwendet, um die gefundenen Ergebnisse der verschiedenen Ansätze und Versuche miteinander zu evaluieren. Dabei wurden implizit die Annahmen getroffen, dass die verwendeten TAQs voneinander unabhängig sind und repräsentativ. Inwiefern die TAQs tatsächlich unabhängig voneinander sind, lässt sich schwer einschätzen. Auch ist unklar, inwiefern die TAQs repräsentativ sind, und somit lassen sich die Ergebnisse nur mit Vorsicht auf weitere TAQs generalisieren. Inwiefern sich die gefundenen Ergebnisse auf Texte außerhalb der Studienordnungen der Universität Bamberg generalisieren lassen, ist ebenfalls unklar. Darüber hinaus wurde keine Power-Analyse durchgeführt, um die Stichprobengröße (hier die Anzahl der TAQs) zu überprüfen, zum Beispiel unter Verwendung der Software G*Power (Faul et al. [2009]). Da die Voraussetzungen für parametrische Tests nicht sicher erfüllt waren, wurde auf die non-parametrischen Tests zurückgegriffen. Zusätzlich sind unterschiedliche Tests unterschiedlich sensibel im Detektieren von möglicherweise signifikanten Ergebnissen (Smucker et al. [2007]). Eine gute Übersicht über die Verwendung von statistischen Tests geben Rainio et al. [2024] und Smucker et al. [2007].

6.3 Generalisierbarkeit über die Anwendung auf Studienordnungen hinaus

Die entwickelten Ansätze sollten auf weitere Studienordnungen der Universität Bamberg sowie auf Studienordnungen anderer Universitäten angewendet werden, um ihre Generalisierbarkeit zu überprüfen. Sofern sich zeigt, dass Studienordnungen in ihrer Formulierung weitgehend ähnlich strukturiert sind, was naheliegend erscheint, kann auch von einer übertragbaren Anwendbarkeit der Ergebnisse ausgegangen werden.

6.4 Offene Fragen und Ausblick

6.4.1 TF-IDF

- **Weitere Sprachen:** Es ist nötig weiter zu testen, ob das Übersetzen von Texten in andere Sprachen als eine Art Pre-Processing bei Verwendung von TF-IDF generell zu besseren Ergebnissen führt. Es kann vermutet werden, dass sich die Anzahl der spezifischen Wörter nach dem Übersetzen unterscheidet: Es könnte vermutet werden, dass Sprachen (bzw. Systeme), die aus einer *low-context* (dies gilt sowohl für Deutschland als auch für die USA) über mehr spezifische Wörter verfügen als solche, die aus einer *high-context* Kultur (bzw. Kommunikationsstil) kommen (vergleiche Gamsriegler [2005]). Im Vergleich zwischen Deutsch und Englisch war die unterschiedliche Anzahl spezifischer Wörter vor allem auf Grund der zusammengesetzten Wörter der Fall. Interessant könnten zum Beispiel Texte des Türkischen sein, da diese Sprache stark agglutinierend ist, jedoch nicht über zusammengesetzte Substantive (als ein Wort) verfügt. Dabei stellen die spezifischen Wörter nur dann ein Problem dar, wenn dadurch eine fehlende Übereinstimmung in Anfrage und Dokumenten entsteht, hier muss weiter geprüft werden, ob dies für alle Kontexte der Fall ist und ob es eine Möglichkeit gibt, dies zu berechnen.
- **Hin- und Rückübersetzung:** Um weiter zu prüfen, wie groß der Einfluss des Übersetzens im Vergleich zum Einfluss der Sprache ist, sollte genauer mit Texten gearbeitet werden, die durch Hin- und Rückübersetzung erstellt wurden. Die Differenz zwischen der Art und Weise, wie Nutzende ihre Informationsbedürfnisse ausdrücken, und dem Inhalt relevanter Dokumente kann dazu führen, dass Dokumente nicht gefunden werden (Wang et al. [2023c]). Die Differenz scheint insbesondere im vorliegenden Kontext, d.h. bei Verwendung von juristischen Texten und „alltäglich formulierten“ Anfragen, der Fall zu sein (vergleiche Abegg and Peric [2021]). Die Differenz im Sprachgebrauch könnte durch Hin- und Rückübersetzung, insbesondere bei Übersetzen in eine Sprache, die wenige spezifische Terme hat (ggf. Sprachen mit *high-context* Kommunikationsstilen), abgeschwächt werden. Um diese Überlegungen zu überprüfen bedarf es weiterer Tests und Recherche. Als Orientierung könnte hier die Arbeit von Saif et al. [2014] dienen, die die Data Sparsity von verschiedenen Textsammlungen berechnet haben.
- **Query Expansion und Document Augmentation:** Query Expansion zeigte eine Verbesserung der Performanz von TF-IDF. Weiter sollte geprüft werden, ob dies auch für Document Augmentation gilt. Außerdem sollte herausgefunden werden, warum die Verbesserung durch Query Expansion insbesondere bei Anfragen auftrat, für die

bereits zumindest mittelmäßige Ergebnisse erzielt wurden, weniger jedoch für Anfragen mit schlechter Performanz. Zum Beispiel geben Azad and Deepak [2019] hierzu einen guten Überblick und Wang et al. [2023b] konnten durch die Verwendung von LLMs zur Query Expansion die Performanz von BM25 um bis zu 15% verbessern. Die Ansätze bringen jedoch auch Risiken mit sich, wie zum Beispiel das Erzeugen vieler irrelevanter Begriffe (Shekarpour et al. [2017]) oder Halluzinationen (Huang et al. [2025]).

- **Optimale Textlänge:** Um die Performanz von TF-IDF weiter zu verbessern, könnte die optimale Textlänge herausgefunden werden. Diese ist vermutlich jedoch stark kontextabhängig und muss auch im Einklang mit den Bedürfnissen der Suchenden stehen. Es zeigte sich, dass besonders kurze Texte (die Titel) sich nicht für die Anwendung von TF-IDF eigneten, denn in sehr kurzen Texten sind relevante Merkmale oft gar nicht vorhanden oder werden nicht erkannt, da TF-IDF den semantischen Kontext nicht erfassen kann (Chen et al. [2013]). Um sicher zu überprüfen, ob TF-IDF tatsächlich eine geringere Performanz in längeren Texten zeigt, reichten die vorhandenen Daten nicht aus, es zeigte sich jedoch eine Tendenz dahingehend. Hier würde sich ein Blick auf BM25 lohnen, da bei BM25 im Gegensatz zu TF-IDF, die Textlänge normalisiert wird (Rosa et al. [2021]).

6.4.2 BERT-basierte Modelle

- **Out-of-Vocabulary (OOV) Detection**: Inwiefern insbesondere spezifische Fachausdrücke nicht im Vokabular von den verschiedenen BERT-Modellen vorhanden waren, und wie dies mit der Performanz zusammenhängt, sollte weiter untersucht werden. Es zeigte sich, dass ML BERT Base auf der ins Englisch übersetzen Studienordnung besser funktionierte, als auf der deutschen Studienordnung [1]. Inwiefern dies mit spezifischeren Begriffen in der deutschen Studienordnung, die nicht im Vokabular vorhanden sind, zusammenhängt könnte durch die Tokenisierungsanalyse, insbesondere auch die Subword Tokenization untersucht werden. Des Weiteren könnte dies auch für die einzelnen Anfragen getestet werden, um so bereits einen Indikator dafür zu haben, wie gut BERT-basierte Modelle für diese Anfrage funktionieren könnte, oder ob ein anderer Ansatz vielversprechender wäre.
- **Merkmale der Anfragen als Indikator für Performanz:** Es scheint Tendenzen dahingehend zu geben, dass Faktfragen besser funktionieren als Listen oder Erweiterte Fragen. Dies könnte jedoch auch ein indirekter Effekt sein, da BERT-basierte Modelle eine bessere Performanz für kurze Textstellen zeigte. Ob nun die Art der Frage oder die Länge der relevanten Textstelle einen größeren Einfluss hat, bzw. ob diese Einflüsse voneinander unabhängig sind, könnte mit größeren Sammlungen an Anfragen und Dokumenten weiter überprüft werden. Es wäre weiter denkbar, dass Faktfragen grundsätzlich von kürzeren Textstellen beantwortet werden und sich daher die Art der Frage als ein Indikator für die Performanz der BERT-basierten Modelle eignen könnte. Dies könnte dann dabei helfen zu entscheiden, ob die Verwendung eines BERT-basierten Modelle eine höhere Performanz erwarten lassen könnte als ein anderer Ansatz oder nicht. Die in dieser Arbeit verwendete Textmenge und vor allem

[1] ML BERT Base unter Verwendung der deutschen Studienordnung: $M = 0{,}33$, $SD = 0{,}45$; ML BERT Base unter Verwendung der ins Englische übersetzten Studienordnung: $M = 0{,}42$, $SD = 0{,}47$ (eigene Metrik, über alle Textstellenebenen)

die kleine Anzahl der Anfragen lässt keine sicheren Schlüsse zu.
- **Modellgröße:** Überraschenderweise zeigte sich eine leicht geringere Performanz in der Anwendung von BERT Large im Vergleiche zu BERT Base (anders als bei GELECTRA Base und Large). Wie es zu diesem Ergebnis gekommen ist bleibt offen.

6.4.3 Mistral Large

- **Sprache:** In der vorliegenden Arbeit wurde Mistral Large lediglich auf der deutschen, originalen Studienordnung getestet. Da jedoch Sprachen im Training der LLMs unterschiedlich stark repräsentiert sind, und dies mit unterschieden in der Performanz einhergeht (Li et al. [2025b]), könnte sich die Anwendung von Mistral Large auf die ins Englische übersetze Texte und Anfragen lohnen.
- **Reihenfolge der Anfragen:** Inwiefern die Reihenfolge der Anfragen und der Textstellen die Antworten von Mistral Large beeinflusst haben, wurde nicht explizit untersucht. Es ist jedoch zu vermuten, dass auch in dem vorliegenden Anwendungskontext ein *Position Bias* auftreten könnte. Beispielsweise untersuchten Zhang et al. [2024] die Möglichkeit, den Einfluss der Reihenfolge durch entsprechendes Prompting zu verringern.
- **Optimierung des Prompts:** Inwiefern die Verbesserung des verwendeten Prompts zu einer Verbesserung der Performanz führen könnte, wurde in dieser Arbeit nicht genauer betrachtet. Beispielsweise Wu et al. [2025] testete die Fähigkeiten von LLMs bei der Extraktion von miRNA-Informationen durch verschiedene Promptstrategien und konnte Verbesserungen in der Performanz erzielen.
- **Typische Fehlerursachen:** Im Rahmen des vorliegenden Anwendungskontexts konnten typische Fehlerursachen identifiziert werden. Inwiefern sich diese Ergebnisse replizieren lassen und auf andere Anwendungskontexte sowie andere LLMs übertragbar sind, müsste weiter geprüft werden.
- **Halluzinationen:** Eine neue Herausforderung in der Anwendung generativer LLMs sind mögliche „Halluzinationen“; hierbei handelt es sich um das Generieren unglaubwürdigen Texts (Peskoff and Stewart [2023]). Dieses Risiko wurde dadurch, dass kein Text generiert werden sollte versucht zu minimieren und trat tatsächlich nicht merkbar auf. Hier fehlt jedoch eine spezifische Auswertung der Texte, ob es im Detail zu Halluzinationen gekommen ist.

6.4.4 Kombinierter Ansatz

- **Zwei-stufiges System: TF-IDF für Feature-Extraction:** Dieses Vorgehen wurde explorativ für ML BERT Base getestet für die ins Englische übersetzten Texte (siehe Unterabschnitt 5.3.2). Hier könnte weiter getestet werden, welcher TF-IDF Wert[2] mit einer guten Performanz einhergeht und wie gut Mistral Large in dieser Anwendung funktioniert.
- **Zwei-stufiges System: TF-IDF als Retriever** TF-IDF könnte, trotz der geringen Textstellenanzahl, als Retriever verwendet werden, um anschließend ML BERT Base als Reranker einzusetzen. Ein solches Vorgehen findet sich beispielsweise bei Nogueira et al. [2019], jedoch nutzen sie BM25 als Retriever. Genauso könnte nach einer ersten

[2]Im durchgeführten Test wurden nur Begriffe mit einem TF-IDF Wert von mindestens 0,30 verwendet.

Anwendung von TF-IDF ein LLM, bzw. Mistral Large, verwendet werden, um die relevante Textstelle, falls vorhanden, herauszufinden.

- **Hybride Modelle:** Beispielsweise kombinierten Aprilio et al. [2025] durch BERT-basierte Modelle und TF-IDF gewonnenen Vektoren, und nutzen diese kombinierten Repräsentationen um ein neuronales Netzwerk für die binäre Klassifikation der Dokumentenrelevanz zu trainieren. Inwiefern dieser Ansatz im gegebenen Kontext anwendbar wäre, müsste geprüft werden. Denkbar wäre eine Gewichtung der Kosinus-Ähnlichkeiten der durch TF-IDF und BERT-basierte Modelle gewonnenen Kosinus-Ähnlichkeiten.
- **Zwei-stufiges System: Topic Modelling:** Da die Themen der TAQs bekannt sind, könnte es vielversprechend sein, beispielsweise mit Hilfe von BERTopic (Grootendorst [2022]), Themen in den Studienordnungen zu identifizieren und zu überprüfen, ob die Auswahl eines Themas als ein erster Schritt (im Sinne des Retrievals) verwendet werden könnte, um anschließend mit einem Reranker die relevante Textstelle zu bestimmen.

6.5 Diskussion der Notwendigkeit des Systems

Aus Umfrage 1 geht hervor, dass es vermutlich zwei Gruppen Studierender gibt: diejenigen, die ihre Studienordnung zu einem großen Teil gelesen haben, und diejenigen, die ihre Studienordnung nur zu einem geringen Teil gelesen haben. Daher ist fraglich, inwiefern ein IR-System, wie es in der vorliegenden Arbeit entworfen wurde, überhaupt genutzt werden würde. Eventuell könnte ein solches IR-System nur von einer der beiden Gruppen überhaupt genutzt werden.

Aus Umfrage 2 kam u. A. die Anmerkung, dass eine Person den Ergebnissen nicht vertrauen würde und die Ergebnisse nochmal selbst überprüfen würde durch eine eigene Suche. Dieser Aspekt wurde bisher in Umfrage 2 nicht erfasst. Damit ein System tatsächlich verwendet wird und den Ergebnissen in einer informierten Art und Weise vertraut werden kann, braucht es Systeme, die die es schaffen, zu erklären, wie es zu den Ergebnissen kommt (Schuetz et al. [2025] und Atf and Lewis [2025]).

Des Weiteren kam in den Freitextrückmeldungen in Umfrage 2 die Frage, warum das System existiert und man nicht einfach über „STRG plus F“ nach der relevanten Textstelle suchen würde. Dieser Aspekt wurde explorativ überprüft, indem die wichtigsten Begriffe in den TAQs in einer Dokumentensuche eingegeben wurden. Dabei wird angenommen, dass es die suchende Person anhand der TAQ bereits unterscheiden kann, ob die Information in der APO oder der StuFPO zu finden ist. Des Weiteren wird angenommen, dass, sofern ein längerer Begriff, häufig ein zusammengesetzter, keine Ergebnisse liefert, der nächstkleinere Teilbegriff betrachtet wird[3].

Faktfrage:

Für die taq_16 „Wie lange ist die Bearbeitungszeit der Abschlussarbeit?“ wurde in der Studien- und Fachprüfungsordnung per „STRG plus F“ mit folgenden Begriffen gesucht:

1. *Bearbeitungszeit* führt sofort zum Erfolg: Der Begriff existiert nur einmal in der Studien- und Fachprüfungsordnung, in der gesuchten Textstelle (anhang_0).

[3]Falls zum Beispiel *Masterarbeit* keine Übereinstimmung liefern würde, würde *Master* verwendet werden

Die Ansätze konnten für die taq_16 folgende Ergebnisse erzielen: Der TF-IDF-Wert für die deutsche Studienordnung beträgt 0,9977, während der Wert für die englische Studienordnung 0,0 ergibt. Für die ins Deutsche rückübersetzte Studienordnung liegt der TF-IDF-Wert bei 0,983. GBERT Base erreicht einen Wert von 0,7745, während das größere GBERT Base Large einen Wert von 0,9977 erzielt, der dem der deutschen Studienordnung im TF-IDF entspricht. Im Gegensatz dazu scheiterten alle Versuche mit Mistral Large.

Für diese TAQ hätte also die Textsuche per „STRG plus F“ schneller zum gewünschten Ergebnis geführt als die verschiedenen Ansätze. TF-IDF unter Verwendung der ins Englische übersetzten Studienordnung hat zwar insgesamt die beste Performanz gezeigt, wäre jedoch für diese TAQ im Vergleich zu den anderen Ansätzen die schlechteste Wahl gewesen. Dies verdeutlicht nochmal, dass ein Indikator dafür, welcher Ansatz am wahrscheinlichsten das richtige Ergebnis liefert, sehr hilfreich wäre.

Listenfrage: Für die taq_19 „Was sind die verschiedenen Prüfungstypen?“ wurde in der APO per „STRG plus F“ mit folgenden Begriffen gesucht:

1. *Prüfungstypen* lieferte keine Übereinstimmung.
2. *Prüfung* lieferte 300 Übereinstimmungen, und eignete sich daher nicht.
3. *verschiedene* lieferte keine Übereinstimmung

Der Begriff, der in einer solchen Suche direkt zum Erfolg geführt hätte, wäre „Prüfungsform“: Dieser Begriff lieferte zwei Übereinstimmungen, beide in dem gesuchten Paragrafen (paragraf_8). Inwiefern eine suchende Person diesen Begriff gewählt hätte, lässt sich ohne weitere Befragung der Nutzenden nicht bewerten.

Die Ansätze konnten für die taq_19 folgende Ergebnisse erzielen: Der TF-IDF-Wert für die deutsche Studienordnung beträgt 0,0, während der Wert bei Verwendung des Titels 0,232 liegt. Für die englische Studienordnung ergibt sich ein Wert von 1,0. Das GBERT Base für die deutsche Sprache erzielt ebenfalls einen Wert von 1,00. Mistral Large zeigte in allen Versuchen korrekte Ergebnisse.

Hier hätten die Ansätze, abgesehen von TF-IDF auf der deutschen Studienordnung, zu einem besseren Ergebnis geführt, als die Suche über „STRG plus F“.

Erweiterte Frage: Für die taq_8 „Wie erhalte ich eine Studienhöchstzeitverlängerung?“ wurde in der APO per „STRG plus F“ mit folgenden Begriffen gesucht:

1. *Studienhöchstzeitverlängerung* lieferte keine Übereinstimmung
2. *Studien* lieferte 75 Übereinstimmungen und eignete sich daher nicht.
3. *Verläng* liefert ein Ergebnis, das jedoch nicht zu dem relevanten Paragrafen führt

Die relevante Textstelle lautet: *Überschreitet die Prüfungskandidatin bzw. der Prüfungskandidat aus nicht von ihr bzw. ihm zu vertretenden Gründen die Ablegungsfrist gemäß Abs. 1, gewährt der Prüfungsausschuss auf Antrag eine Nachfrist.*

Die Ansätze konnten für die taq_8 folgende Ergebnisse erzielen: Der TF-IDF-Wert für die englische Studienordnung beträgt 0,983, während der Wert für die deutsche Studienordnung 0 beträgt. Das BERT-Modell mit Query Expansion erzielt einen Wert von 0,983. Im Gegensatz dazu lieferte das Mistral-Modell durchmischte Ergebnisse.

Daraus lässt sich vermuten, dass die verwendeten Ansätze tatsächlich einen Vorteil für die Suchenden liefern. Außerdem untermauern diese Beispiele nochmal, dass Studierende einen anderen Sprachgebrauch haben als ihre Studienordnungen.

Implikationen:

1. Damit Suchende Vertrauen in die Ergebnisse haben können, könnte es hilfreich sein, Ansätze zu wählen, die sich gut erklären lassen. Dies spricht für die Verwendung von TF-IDF.

2. Um Suchenden zu helfen abschätzen zu können, wie vertrauenswürdig die Ergebnisse sind, könnte geprüft werden, ob sich dies aus der Varianz der Kosinus-Ähnlichkeiten der verwendeten Ansätze ableiten lässt.

3. Die Entscheidung für den für die spezifische TAQ am besten funktionierenden Ansätze anhand der Merkmale der TAQ richtig zu treffen könnte die Performanz des Systems nochmal steigern. Hier würde sich insbesondere lohnen genauer zu betrachten, unter welchen Bedingungen die deutschen TAQs und Studienordnungen zu einem besseren Ergebnis führen als die Verwendung der ins Englische übersetzten TAQs und Studienordnungen.

6.6 Abschließende Bewertung anhand der Leitfragen

1.1 Welchen Inhalt hat das Informationsbedürfnis der Zielgruppe im Kontext von Studien- und Fachprüfungsordnungen?

Die Frage konnte durch die gefundenen TAQs größtenteils beantwortet werden, jedoch waren nicht alle Gruppen in der dafür verwendeten Umfrage vertreten. Insbesondere scheinen Studierende in ihrem Informationsbedürfnis zum Teil keinen Unterschied zwischen der Studienordnung und den Modulhandbüchern zu machen, was für die Entwicklung eines übergreifenden Systems spricht.

1.2 Mit welcher Performanz eines möglichen Informationssystems wäre die Zielgruppe zufrieden?

Die Umfrage 2 ergab ein gut interpretierbares Bild dessen, mit welcher Performanz eines möglichen Informationssystems die Zielgruppe zufrieden wäre. Eine genauere Betrachtung z.B. durch Prototypentests könnte die Ergebnisse aus Umfrage 2 weiter ausbauen.

1.3 Wie kann ein Informationssystem für die Zielgruppe aufgebaut sein?

In dieser Arbeit wurde sich für einen möglichen Aufbau eines Systems ohne Unterscheidung zwischen *Retriever* und *Reranker* entschieden. Diese Entscheidung erfolgte aufgrund der vergleichsweise geringen Anzahl potenzieller Treffer. Ob dieser Ansatz zu einer besseren Performanz führt und wie er im Vergleich zu alternativen Systemarchitekturen abschneidet, konnte im Rahmen dieser Arbeit nicht beantwortet werden.

2.1 Welche Ansätze kommen für den gegebenen Kontext in Frage?

Es zeigt sich eine klare Tendenz dahingehend, dass TF-IDF in Kombination mit dem Übersetzen der Texte in die englische Sprache ein vielversprechender Ansatz ist. Jedoch konnte die Generalisierbarkeit auf andere Studienordnungen zwar vermutet, jedoch nicht getestet werden. Weitere Aspekte für die Entscheidung könnten z.B. die Erklärbarkeit der verwendeten Ansätze sowie der Ressourcenverbrauch und die Determiniertheit sein.
Wie kann die Performanz der verwendeten Ansätze evaluiert werden?

2.2.1 Hinsichtlich der Vergleichbarkeit mit anderen IR-Systemen?

Die Verwendung einer spezifischen, auf Umfrage 2 basierenden Metrik schränkte die Vergleichbarkeit mit anderen Systemen ein. Ergänzend wurden jedoch auch etablierte Metriken wie der Reciprocal Rank berechnet, um eine bessere Vergleichbarkeit zu ermöglichen. Darüber hinaus wurde die Nutzerzufriedenheit als qualitatives Bewertungskriterium herangezogen.

2.2.2 Hinsichtlich der Passung zu den Bedürfnissen der Zielgruppe?

Durch die auf Umfrage 2 aufbauende Metrik ermöglichte eine erste Bewertung der Passung zur Zielgruppe, jedoch einzig bezogen auf den Rang, in dem eine relevante Textstelle gefunden wurde. Weitere Kriterien, die für die Passung des Systems zur Zielgruppe relevant sind, wurden in dieser Arbeit nur in geringem Maß berücksichtigt.

2.2.3 Können bestimmte Eigenschaften der Anfragen genutzt werden, um die Performanz der Ansätze vorherzusagen und dadurch eine effektive Kombination zu ermöglichen?

Zwar konnten erste Indizien identifiziert werden, doch aufgrund der geringen Stichprobengröße sind weitere Tests erforderlich. Der Versuch, mithilfe der Varianz der Kosinus-Ähnlichkeiten auf Ebene einzelner Anfragen eine Vorhersage zu treffen, blieb erfolglos. Allerdings zeigte sich, dass der Mittelwert der Varianzen aller Anfragen bei Anwendungen, die eine insgesamt bessere Performanz aufwiesen, tendenziell höher war, als bei Anwendungen mit einer geringeren Performanz. Inwiefern dieses Ergebnis generalisierbar ist, muss jedoch weiter geprüft werden.

Literaturverzeichnis

Abegg, A. and Peric, B. (2021). *Sprache und Sprachgebrauch des Rechts: eine korpuslinguistische Diskursanalyse auf Basis der Entscheide des schweizerischen Bundesgerichts und der Botschaften des Bundesrats*. Dike.

Abubakar, H. D., Umar, M., and Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. *SLU Journal of Science and Technology*, 4(1):27–33.

Allam, A. M. N. and Haggag, M. H. (2012). The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Allen, K., Berry, M. M., Luehrs Jr, F. U., and Perry, J. W. (1955). Machine literature searching viii. operational criteria for designing information retrieval systems. *American Documentation (pre-1986)*, 6(2):93.

Appenzeller, P. and Kersting, R. (2013). *Endlich Studium!: Das Handbuch für die beste Zeit deines Lebens*. rap Verlag.

Aprilio, P., Felix, M., Nugraha, P. S., and Fahmi, H. (2025). Hybrid feature combination of tf-idf and bert for enhanced information retrieval accuracy. *JISA (Jurnal Informatika dan Sains)*, 8(1):8–15.

Arora, M., Kanjilal, U., and Varshney, D. (2016). Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, 12(2):224–236.

Assal, H., Seng, J., Kurfess, F., Schwarz, E., and Pohl, K. (2011). Semantically-enhanced information extraction. In *2011 Aerospace Conference*, pages 1–14. IEEE.

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4):340–350.

Atf, Z. and Lewis, P. R. (2025). Is trust correlated with explainability in ai? a meta-analysis. *IEEE Transactions on Technology and Society*.

Aula, A. (2003). Query formulation in web information search. In *ICWI*, pages 403–410.

Azad, H. K. and Deepak, A. (2019). Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, 56(5):1698–1735.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Bama, S. S., Ahmed, M. S. I., and Saravanan, A. (2018). A SURVEY ON PERFORMANCE EVALUATION MEASURES FOR INFORMATION RETRIEVAL SYSTEM. *International Research Journal of Engineering and Technology (IRJET)*, 02(02).

Becker, A. (2001). Interdisziplinäre arbeitsgruppe „sprache des rechtsän der berlin-brandenburgischen akademie der wissenschaften-projektdarstellung. In *Sprache und Recht*, pages 361–365. 365.

Belkin, N. J. (2000). Helping people find what they don't know. *Communications of the ACM*, 43(8):58–61.

Belkin, N. J. et al. (1993). Interaction with texts: Information retrieval as information seeking behavior. *Information retrieval*, 93(55-66).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bush, V. et al. (1945). As we may think. *The atlantic monthly*, 176(1):101–108.

Cao, Y.-g., Cimino, J. J., Ely, J., and Yu, H. (2010). Automatically extracting information needs from complex clinical questions. *Journal of biomedical informatics*, 43(6):962–971.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. *arXiv preprint arXiv:2010.10906*.

Chapelle, O., Metlzer, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630.

Chen, M., Weinberger, K. Q., Sha, F., et al. (2013). An alternative text representation to tf-idf and bag-of-words. *arXiv preprint arXiv:1301.6770*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Cosijn, E. and Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management*, 36(4):533–550.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Dejprapatsorn, P., Uypatchawong, S., and Songmuang, P. (2025). Bert-based semantic retrieval for academic abstracts. In *2025 IEEE International Conference on Cybernetics and Innovations (ICCI)*, pages 1–6. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ding, Z., Huang, Y., Yuan, H., and Dong, H. (2020). Introduction to reinforcement learning. *Deep reinforcement learning: fundamentals, research and applications*, pages 47–123.

Djeddal, H., Erbacher, P., Toukal, R., Soulier, L., Pinel-Sauvagnat, K., Katrenko, S., and Tamine, L. (2024). An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5354–5359.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160.

Fidel, R. (1993). Qualitative methods in information retrieval research. *Library and information science research*, 15:219–219.

Fu, X., Yilmaz, E., and Lipani, A. (2022). Evaluating the cranfield paradigm for conversational search systems. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 275–280.

Gamsriegler, A. (2005). High-context and low-context communication styles. *Studiengang Informationsberufe*, pages 1–8.

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., and Alshemaimri, B. (2025). Bert applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6):1–49.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, pages 135–140. Elsevier.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., and Cheng, X. (2022). Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42.

Guo, T., Guo, J., Fan, Y., Lan, Y., Xu, J., and Cheng, X. (2018). A comparison between term-based and embedding-based methods for initial retrieval. In *China Conference on Information Retrieval*, pages 28–40. Springer.

Hambarde, K. A. and Proença, H. (2023). Information Retrieval: Recent Advances and Beyond. *IEEE Access*, 11:76581–76604.

Heie, M. H., Whittaker, E. W., and Furui, S. (2012). Question answering using statistical language modelling. *Computer Speech & Language*, 26(3):193–209.

Helmchen, J. F. (2017). Verständliche rechtssprache-ein steiniger weg/eingereicht von joachim helmchen.

Herbst, U., Voeth, M., Eidhoff, A. T., Müller, M., and Stief, S. (2016). Studierendenstress in Deutschland – eine empirische Untersuchung.

Hilgers, R.-D., Heussen, N., and Stanzel, S. (2018). Nullhypothese. In *Lexikon der Medizinischen Laboratoriumsdiagnostik*, pages 1–1. Springer.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hirschman, L. and Gaizauskas, R. (2001). Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550.

Hofmann, K., Li, L., Radlinski, F., et al. (2016). Online evaluation for information retrieval. *Foundations and Trends® in Information Retrieval*, 10(1):1–117.

Hoskin, T. (2012). Parametric and nonparametric: Demystifying the terms. In *Mayo Clinic*, volume 5, pages 1–5.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ingwersen, P. (1992). *Information retrieval interaction*, volume 246. Taylor Graham London.

Jiang, P. and Cai, X. (2024). A Survey of Text-Matching Techniques. *Information*, 15(6):332.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum*, volume 51, pages 4–11. Acm New York, NY, USA.

Kadhim, A. I. (2019). Term weighting for feature extraction on twitter: A comparison between bm25 and tf-idf. In *2019 international conference on advanced science and engineering (ICOASE)*, pages 124–128. IEEE.

Kagolovsky, Y. and Möhr, J. R. (2001). A new approach to the concept of "relevance" in information retrieval (ir). In *MEDINFO 2001*, pages 348–352. IOS Press.

Kamaluddin, M. I., Rasyid, M. W. K., Abqoriyyah, F. H., and Saehu, A. (2024). Accuracy analysis of deepl: Breakthroughs in machine translation technology. In *Journal of English Education Forum (JEEF)*, volume 4, pages 122–126.

Kang, H. (2021). Sample size determination and power analysis using the g* power software. *Journal of educational evaluation for health professions*, 18.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs].

Kekäläinen, J. (2005). Binary and graded relevance in ir evaluations—comparison of the effects on ranking of ir systems. *Information processing & management*, 41(5):1019–1033.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., and Dou, Z. (2025a). From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Li, Z., Shi, Y., Liu, Z., Yang, F., Payani, A., Liu, N., and Du, M. (2025b). Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., Zhang, L., Li, Z., and Ma, Y. (2024). Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luan, Y., Eisenstein, J., Toutanova, K., and Collins, M. (2021). Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Mandikal, P. and Mooney, R. (2024). Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval. arXiv:2401.04055 [cs].

Mandl, T. (2010). Evaluierung im information retrieval. *Information–Wissenschaft und Praxis*, 61.

Marwah, D. and Beel, J. (2020). Term-recency for tf-idf, bm25 and use term weighting. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 36–41.

McDonald, D., Papadopoulos, R., and Benningfield, L. (2024). Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Preprints*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Mistral AI (2024). Large enough. Abgerufen am 6. September 2025.

Mitra, B. and Craswell, N. (2018). An Introduction to Neural Information Retrieval t. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

Moffat, A. and Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27.

Moghadasi, S. I., Ravana, S. D., and Raman, S. N. (2013). Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics*, 7(2):301–312.

Neves, M. and Leser, U. (2015). Question answering for biology. *Methods*, 74:36–46.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Nogueira, R., Yang, W., Cho, K., and Lin, J. (2019). Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Penha, G., Câmara, A., and Hauff, C. (2022). Evaluating the robustness of retrieval pipelines with query variation generators. In *European conference on information retrieval*, pages 397–412. Springer.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Perumal, T., Mustapha, N., Mohamed, R., and Shiri, F. M. (2024). A Comprehensive Overview and Comparative Analysis on Deep Learning Models. *Journal on Artificial Intelligence*, 6(1):301–360.

Peskoff, D. and Stewart, B. M. (2023). Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438.

Pudaruth, S., Boodhoo, K., and Goolbudun, L. (2016). An intelligent question answering system for ict. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 2895–2899. IEEE.

Qaiser, S. and Ali, R. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 181(1):25–29.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *arXiv preprint arXiv:1810.04805*.

Radlinski, F. and Craswell, N. (2013). Optimized interleaving for online retrieval evaluation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 245–254.

Radlinski, F., Kurup, M., and Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 43–52.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Raghavan, V. V. and Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science*, 37(5):279–287.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Rice, M. E. and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: Roc area, cohen's d, and r. *Law and human behavior*, 29(5):615–620.

Robertson, S. E. and Hancock-Beaulieu, M. M. (1992). On the evaluation of ir systems. *Information Processing & Management*, 28(4):457–466.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). *Okapi at TREC-3*. British Library Research and Development Department.

Rosa, G. M., Rodrigues, R. C., Lotufo, R., and Nogueira, R. (2021). Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of twitter.

Salton, G. (1975). A vector space model for information retrieval. *Journal of the ASIS*, pages 613–620.

Saracevic, T. (1995). Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–146.

Schuetz, S., Kuai, L., Lacity, M. C., and Steelman, Z. (2025). A qualitative systematic review of trust in technology. *Journal of Information Technology*, 40(1):55–76.

Sendatzki, S. and Rathmann, K. (2022). Unterschiede im Stresserleben von Studierenden und Zusammenhänge mit der Gesundheit. Ergebnisse einer Pfadanalyse. *Prävention und Gesundheitsförderung*, 17(4):416–427.

Shekarpour, S., Marx, E., Auer, S., and Sheth, A. (2017). Rquery: rewriting natural language queries on knowledge graphs to alleviate the vocabulary mismatch problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE.

Singh, S. (2018). Natural language processing for information extraction. *arXiv preprint arXiv:1807.02383*.

Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Smucker, M. D., Allan, J., and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632.

Soares, M. A. C. and Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management*, 31(3):397–417.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsai, H.-C., Huang, Y.-F., and Kuo, C.-W. (2024). Comparative analysis of automatic literature review using mistral large language model and human reviewers.

Vasileiou, A. and Eberle, O. (2024). Explaining text similarity in transformer models. *arXiv preprint arXiv:2405.06604*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is All you Need. *arXiv preprint arXiv:1706.03762*.

Voorhees, E. M. (2019). The evolution of cranfield. In *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, pages 45–69. Springer.

Wang, H., Li, J., Wu, H., Hovy, E., and Sun, Y. (2023a). Pre-trained language models and their applications. *Engineering*, 25:51–65.

Wang, L., Yang, N., and Wei, F. (2023b). Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Wang, X., MacAvaney, S., Macdonald, C., and Ounis, I. (2023c). Generative query reformulation for effective adhoc search. *arXiv preprint arXiv:2308.00415*.

Wicaksono, A. F. and Moffat, A. (2020). Metrics, User Models, and Satisfaction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 654–662, Houston TX USA. ACM.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

Wu, R., Zong, H., Wu, E., Li, J., Zhou, Y., Zhang, C., Zhang, Y., Wang, J., Tang, T., and Shen, B. (2025). Improving large language models for mirna information extraction via prompt engineering. *Computer Methods and Programs in Biomedicine*, page 109033.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yunanda, G., Nurjanah, D., and Meliana, S. (2022). Recommendation system from microsoft news data using tf-idf and cosine similarity methods. *Building of informatics, technology and science (BITS)*, 4(1):277–284.

Zhang, M., Meng, Z., and Collier, N. (2024). Can we instruct llms to compensate for position bias? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12545–12556.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2024). Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J.-R. (2023). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J.-R. (2024). Large Language Models for Information Retrieval: A Survey. arXiv:2308.07107 [cs].

Zhuang, S., Ma, X., Koopman, B., Lin, J., and Zuccon, G. (2024). Promptreps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. *arXiv preprint arXiv:2404.18424*.

Zuva, K. and Zuva, T. (2012). Evaluation of information retrieval systems. *International journal of computer science & information technology*, 4(3):35.

Erklärung

Ich erkläre hiermit gemäß § 9 Abs. 12 APO, dass ich die vorstehende Masterarbeit selbstständig verfasst bzw. erbracht habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt worden sind. Ferner, dass die digitale Fassung der gedruckten Ausfertigung ausnahmslos in Inhalt und Wortlaut entspricht und dass zur Kenntnis genommen wurde, dass die digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Datum	Unterschrift

Kapitel 7

Anhang

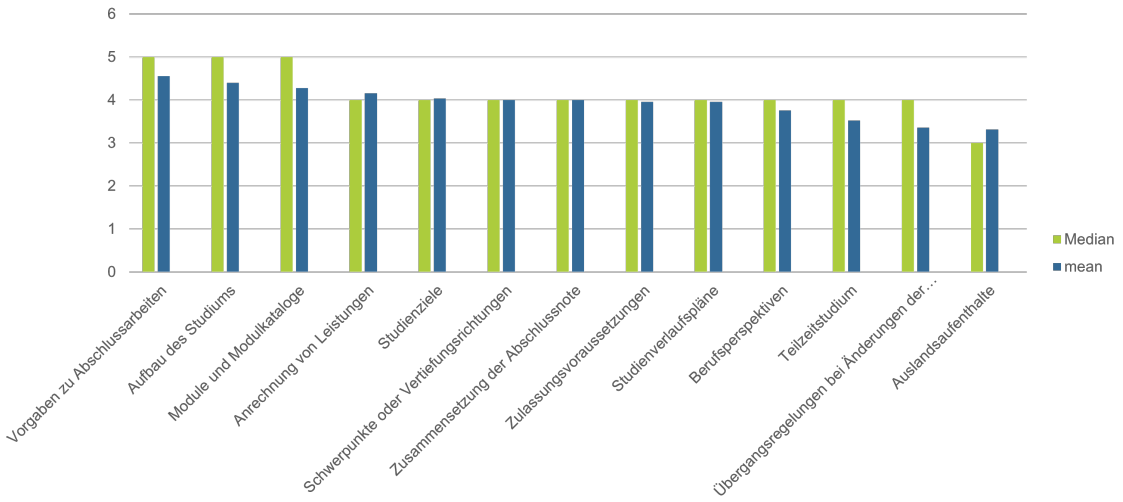

Abbildung 13: *Ergebnisse aus Umfrage 1: Gewichtung verschiedener Themen nach Wichtigkeit durch die Teilnehmenden, auf einer Skala von 0 (sehr irrelevant) bis 5 (sehr wichtig).*

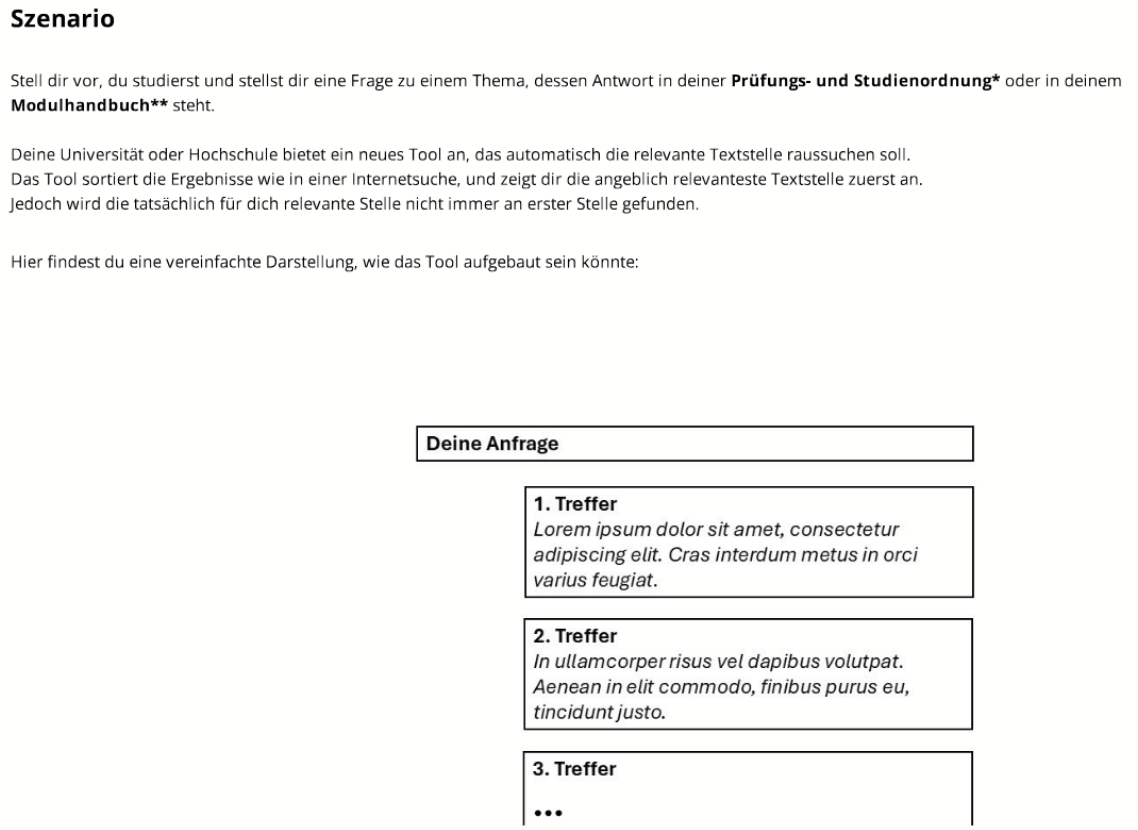

Abbildung 14: *Ausschnitt aus der Umfrage 2: Szenario und Fragen zur Bewertung der erwarteten Performanz eines IR-Systems zur Unterstützung bei studienbezogenen Informationsanfragen zu Studienordnung oder Modulhandbüchern*

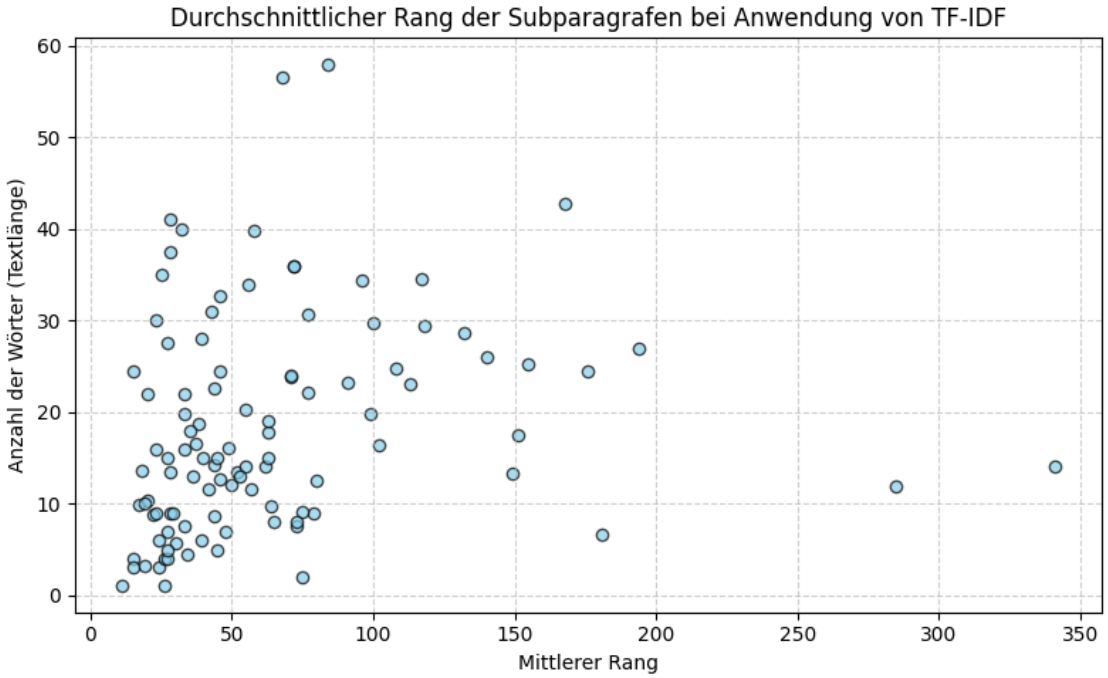

Abbildung 15: *Durchschnittlicher Rang der Subparagrafen nach Textlänge bei Anwendung von TF-IDF, nach Ausschluss von Kosinus-Ähnlichkeiten mit dem Wert null (Anzahl der Wörter vor dem Preprocessing).*

Abbildung 16: *Word Cloud für die basierend auf der deutschen Studienordnung.*

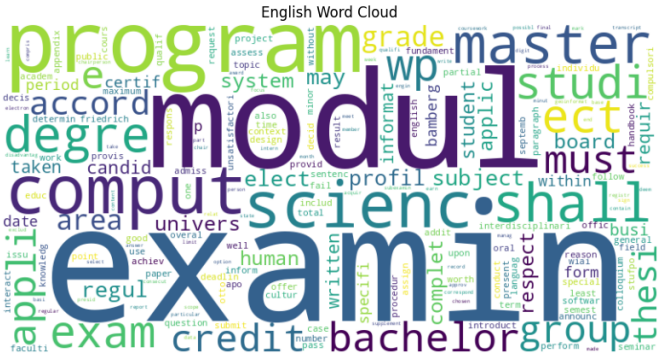

Abbildung 17: *Word Cloud für die basierend auf der ins Englische übersetzten Studienordnung.*

Abbildung 18: *Word Cloud für die basierend auf der vom Englischen ins Deutsche rückübersetzten Studienordnung.*

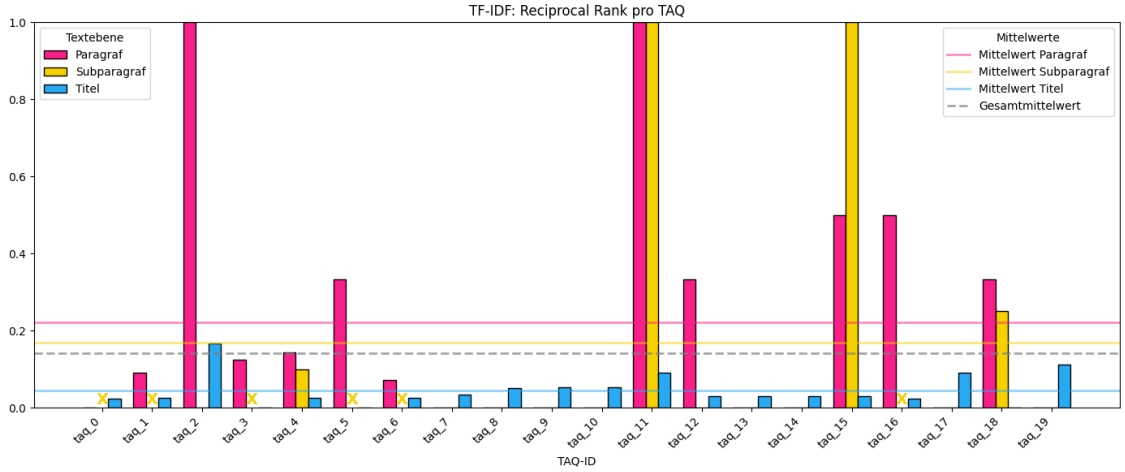

Abbildung 19: *Reciprocal Rank Werte jede TAQ basierend auf TF-IDF unter Verwendung der originalen Studienordnung.*

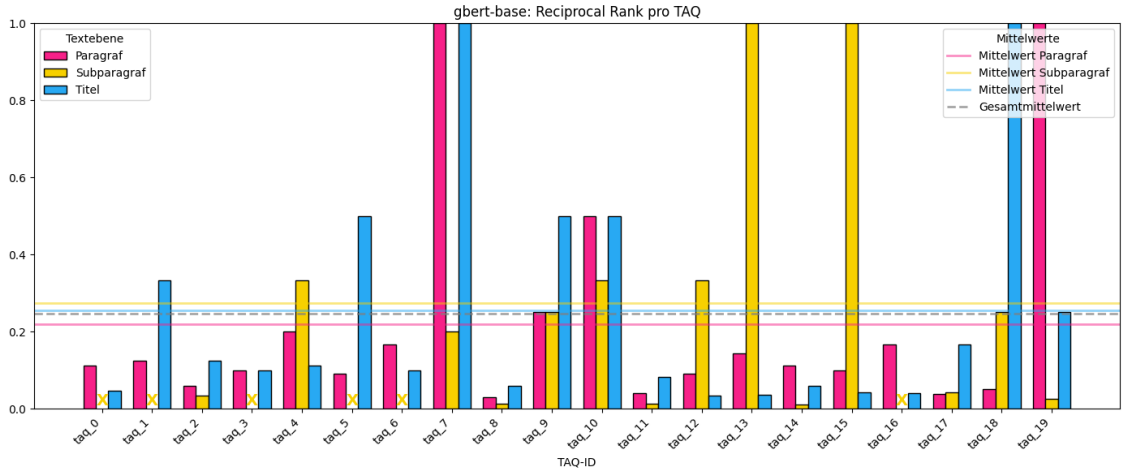

Abbildung 20: *Reciprocal Rank Werte jede TAQ basierend auf gbert-base unter Verwendung der originalen Studienordnung.*

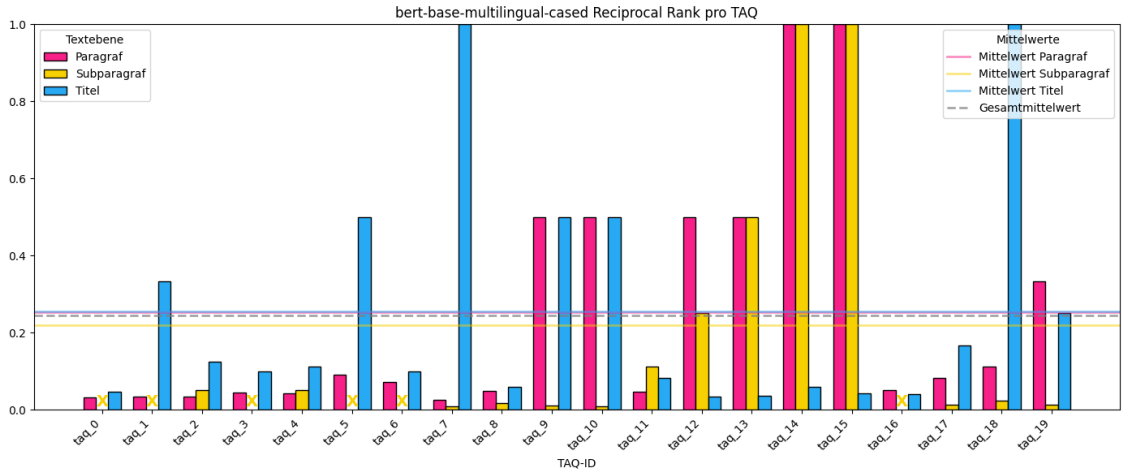

Abbildung 21: *Reciprocal Rank Werte jede TAQ basierend auf ML BERT Base unter Verwendung der originalen Studienordnung.*

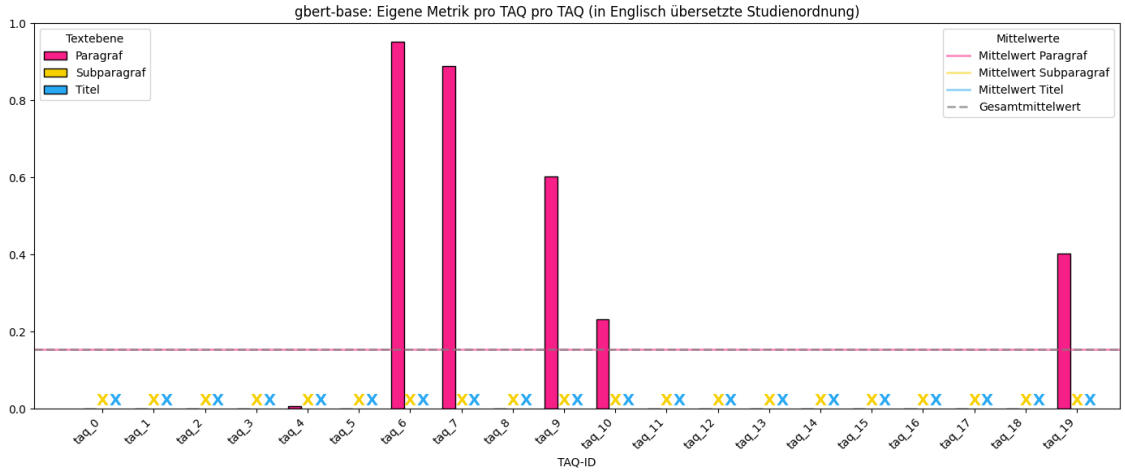

Abbildung 22: *Werte der eigenen Metrik für jede TAQ basierend auf GBERT unter Verwendung der ins Englische übersetzten Studienordnung.*

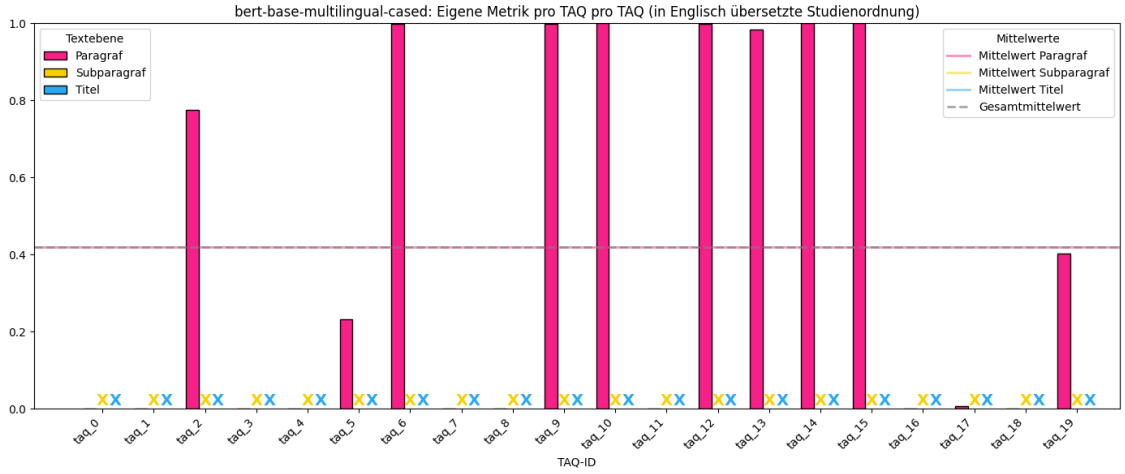

Abbildung 23: *Werte der eigenen Metrik für jede TAQ basierend auf BERT_ML unter Verwendung der ins Englische übersetzten Studienordnung.*

Profil			ID	Modulbezeichnung	ECTS	Prüfung
1	2	3				
Modulgruppe A1 – Pflichtbereich: Profil 1 45 ECTS, Profil 2 27 – 36 ECTS, Profil 3 15 – 27 ECTS						
P			Kinf-IPKult-E	Informatik und Programmierung für die Kultur-wissenschaften	9	Hausarbeit und Klausur
P	P	P	Inf-DM-B	Diskrete Modellierung	9	Klausur
P	E		Inf-Einf-B	Einführung in die Informatik	9	Klausur
P	P	P	SWT-FSE-B	Foundations of Software Engineering	6	Klausur
P	P	E	AI-AuD-B	Algorithmen und Datenstrukturen	6	Klausur
P	P	E	MOBI-DBS-B	Datenbanksysteme	6	Klausur
Modulgruppe A1 – Wahlpflichtbereich: 6 ECTS						
WP	WP	WP	WiMa-B-001	Wirtschaftsmathematik: Lineare Algebra	6	Klausur
WP	WP	WP	WiMa-B-002	Wirtschaftsmathematik: Analysis	6	Klausur

Abbildung 24: *Beispiel der Darstellung der wählbaren Modulgruppen aus dem Anhang der Studienordnung.*

Tabelle 21: *Übersicht der Typical Asked Questions (TAQs) samt Thema der TAQ*

TAQ ID	**Thema**	**Frage**
taq_0	Inhalt des Studiums	Welche Fächer kann ich wählen?
taq_1	Inhalt des Studiums	Ist nur eine bestimmte Kombination in der Fächerwahl zulässig?
taq_2	Inhalt des Studiums	Welche Fächer kann ich mir anrechnen lassen?
taq_3	Inhalt des Studiums	Welche Module sind im Wahlbereich belegbar?
taq_4	Inhalt des Studiums	Unter welchen Bedingungen kann ich Fächer außerhalb des Modulhandbuches wählen?
taq_5	Inhalt des Studiums	Welche Module sind in den Modulgruppen wählbar?
taq_6	Inhalt des Studiums	Kann ich Pflichtmodule durch andere Module ersetzen?
taq_7	Dauer des Studiums	Wie lange kann ich studieren, bis ich exmatrikuliert werde?
taq_8	Dauer des Studiums	Wie erhalte ich eine Studienhöchstzeitverlängerung?
taq_9	Dauer des Studiums	Muss ich die insgesamt 180 ECTS bzw. 120 ECTS genau erreichen?
taq_10	Dauer des Studiums	Was passiert, wenn ich mehr als 180 ECTS bzw. 120 ECTS erreiche?
taq_11	Dauer des Studiums	Was hat es mit der Studienfortschrittskontrolle auf sich?
taq_12	Abschlussarbeit	Was sind die inhaltlichen Erwartungen an eine Abschlussarbeit?
taq_13	Abschlussarbeit	Wie viele ECTS müssen erreicht werden, um die Abschlussarbeit anmelden zu können?
taq_14	Abschlussarbeit	An welchem Lehrstuhl kann ich meine Abschlussarbeit schreiben?
taq_15	Abschlussarbeit	Wie ist die Gewichtung der Bewertung der Abschlussarbeit?
taq_16	Abschlussarbeit	Wie lange ist die Bearbeitungszeit der Abschlussarbeit?
taq_17	Prüfungen	Wie oft darf man durchfallen?
taq_18	Prüfungen	Kann man Prüfungen wiederholen, um seine Note aufzubessern?
taq_19	Prüfungen	Was sind die verschiedenen Prüfungstypen?

Tabelle 22: *Deskriptive Kennwerte (Mittelwerte und Standardabweichungen) der Items zur akzeptierten Position einer relevanten Textstelle im IR-System, differenziert nach empfundener Nützlichkeit, Weiterempfehlung und Zufriedenheit.*

Item	**Frage**	**M**	**SD**
Zufrieden	Bitte gebe die maximale Stelle an, an der die relevante Stelle stehen darf, damit du mit dem Tool zufrieden bist.	3,65	2,16
Weder zufrieden noch unzufrieden	Bitte gebe die durchschnittliche Stelle an, an der die relevante Stelle stehen muss, damit du mit dem Tool weder zufrieden noch unzufrieden bist.	5,62	4,51
Unzufrieden	Bitte gebe die erste Stelle an, an der die relevante Stelle stehen muss, damit du mit dem Tool unzufrieden bist.	8,19	6,73
Nützlich	Bitte gebe die maximale Stelle an, an der die relevante Stelle stehen darf, damit das Tool nützlich ist.	4,96	3,07
Weder nützlich noch nutzlos	Bitte gebe die durchschnittliche Stelle an, an der die relevante Stelle stehen muss, damit das Tool weder nützlich noch nutzlos ist.	7,37	4,21
Nutzlos	Bitte gebe die erste Stelle an, an der die relevante Stelle stehen muss, damit das Tool nutzlos ist.	11,56	7,75
Weiterempfehlen	An welcher Stelle darf die relevante Textstelle maximal stehen, damit du das Tool weiterempfehlen würdest?	4,58	2,71

Tabelle 23: *Darstellung der Wortanzahl vor und nach dem Pre-Processing und der Häufigkeit der Relevanz für eine TAQ aller Paragrafen*

Paragraf_ID	**Anzahl der Wörter vor dem Preprocessing (TF-IDF)**	**Anzahl der Wörter nach dem PreProcessing (TF-IDF)**	**Häufigkeit relevant**
paragraf_0	2478	171	0
paragraf_1	732	54	0
paragraf_2	2301	185	0
paragraf_3	4733	307	0
paragraf_4	926	66	0
paragraf_5	1992	127	1
paragraf_6	3002	198	1
paragraf_7	846	53	0
paragraf_8	12547	823	1
paragraf_9	2729	172	0
paragraf_10	2216	143	3
paragraf_11	570	38	0
paragraf_12	1147	71	0
paragraf_13	267	13	0
paragraf_14	1275	88	0
paragraf_15	1412	95	0
paragraf_16	1749	123	0
paragraf_17	407	26	0
paragraf_18	370	22	2
paragraf_19	1667	106	1
paragraf_20	3846	267	1
paragraf_21	595	34	0
paragraf_22	611	37	0
paragraf_23	970	64	0
paragraf_24	1328	84	0
paragraf_25	885	59	0
paragraf_26	185	12	0
paragraf_27	1772	114	0
paragraf_28	565	36	1
paragraf_29	200	14	1
paragraf_30	508	35	0
paragraf_31	657	46	0
paragraf_32	1056	70	0
paragraf_33	1024	63	0
paragraf_34	1221	79	4
paragraf_35	433	34	0
paragraf_36	1825	124	0
paragraf_37	259	16	0
paragraf_38	2019	139	0
paragraf_39	1945	142	0
paragraf_40	11332	898	3
anhang_0	8356	700	4
anhang_1	2036	135	0

Tabelle 24: *Beispielhafte Übersetzung einer relevanten Textstelle aus der Studienordnung*

Sprache	**Relevante Textstelle aus anhang_0**
Deutsch	Die Modulprüfung wird durch schriftliche Hausarbeit mit einer Bearbeitungszeit von sechs Monaten und einem Kolloquium mit einer Prüfungsdauer von 20 bis 60 Minuten erbracht.
Englisch	The module examination is completed by means of a written assignment with a processing time of six months and a colloquium lasting 20 to 60 minutes.

Tabelle 25: *Wörter mit TF-IDF Werten über 0.3 per Paragraph (Englisch und Deutsch)*

Paragraph	**Englische Übersetzung**	**Deutsches Original**
paragraph_0	program, science	geltungsbereich
paragraph_1	ect, modul, point	prufung
paragraph_2	degree, science, system	bachelor, mastergrad
paragraph_3	board, chairperson, examin	prufungsausschuss
paragraph_4	appoint, examin, propos	beisitz, beisitzerinn, pruf, pruferinn
paragraph_5	achiev, credit	anerkenn, praktikumsleist, prufungsleist, studienzeit
paragraph_6	examin, withdraw	ordnungsverstoss, rucktritt, tauschung, versaumnis
paragraph_7	defici, examin	mangel, prufungsverfahr
paragraph_8	examin	durchfuhr, form, prufung
paragraph_9	grade, perform	bewertet, prufungsleist
paragraph_10	examin, fail, modul	besteh, modul, modulpruf, wiederhol
paragraph_11	cours, held, modul	lehrveranstalt
paragraph_12	examin, registr	anmeldetermin, prufung
paragraph_13	admiss, enrol	zugang, zulassungsvoraussetz
paragraph_14	modul, registr	meldefrist, zulassungsverfahr
paragraph_15	thesi, topic	bachelor, masterarbeit, zulass
paragraph_16	bachelor, thesi	bachelor, bewertet, form, masterarbeit
paragraph_17	bachelor, master, repetit, second, thesi	bachelor, masterarbeit, wiederhol
paragraph_18	complet, degree, modul, success	abschluss, erfolgreich, studiengang
paragraph_19	definit, examin, fail	endgult, nichtbesteh
paragraph_20	certif	urkund, zeugnis
paragraph_21	addit	zusatzpruf
paragraph_22	compens, disabl, disadvantag	behindert, nachteilsausgleich
paragraph_23	—	gesetz, nachteilsausgleich, schutzbestimm, schwang
paragraph_24	examin	prufung, ungult
paragraph_25	examin, inspect	einsicht, prufungsakt
paragraph_26	announc, locat, public	bekanntmach, offent
paragraph_27	bamberg, friedrich, march, otto	inkrafttret
paragraph_28	informat, program, regul	geltungsbereich
paragraph_29	period, scope, semest, studi	studiendau, studienumfang
paragraph_30	comput, degree, informat, relat	studiengang, verwandt
paragraph_31	handbook, modul	modulhandbuch
paragraph_32	second	zugangsvoraussetz
paragraph_33	—	gegenstand, masterstudiengang
paragraph_34	colloquium, thesi	masterarbeit
paragraph_35	focus, indic	studienschwerpunkt
paragraph_36	comput, profil, science	studiengangsprofil
paragraph_37	knowledg	studienvoraussetz
paragraph_38	—	studium, ziel
paragraph_39	comput, group, science	struktur, studium
paragraph_40	modul, wp	ausserkrafttret, inkrafttret, ubergangsregel
paragraph_41	modul, wp	computing, humaniti, masterstudiengang, modul, modulgrupp, the
paragraph_42	—	computing, humaniti, masterarbeit, masterstudiengang, the, themengebiet

Tabelle 26: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch GBERT Base berechneten Ergebnisse für die eigene Metrik unter Verwendung der ins Englische übersetzten Studienordnung.*

Category	**Mean**	**SD**	**Median**
Overall	0,15	0,31	0,00
Textebene			
Paragraphs	0,15	0,31	0,00
Subparagraphs	–	–	–
Titles	–	–	–
Fragenart			
Erweiterte Frage	0,04	0,09	0,00
Faktfrage	0,31	0,43	0,00
Listenfrage	0,07	0,16	0,00

Tabelle 27: *Mittelwerte (M), Standardabweichungen (SD) und Mediane (MD) der durch ML BERT Base berechneten Ergebnisse für die eigene Metrik unter Verwendung der ins Englische übersetzten Studienordnung.*

Category	**Mean**	**SD**	**Median**
Overall	0,42	0,47	0,12
Textebene			
Paragraphs	0,42	0,47	0,12
Subparagraphs	–	–	–
Titles	–	–	–
Fragenart			
Erweiterte Frage	0,33	0,52	0,00
Faktfrage	0,50	0,53	0,49
Listenfrage	0,40	0,41	0,32

Tabelle 28: *Vermutete Fehlerursachen in der Verwendung von Mistral Large im Versuch 1. Betratet werden nur die TAQs, bei denen die Antwort von Mistral Large in den zwei Durchläufen übereinstimmten.*

taq_id	**Textebene**	**Fehlerursache**
taq_2	Paragraf	Falscher Kontext
taq_6	Paragraf	Antwort nur implizit enthalten
taq_7	Paragraf	Falscher Kontext
taq_16	Paragraf	Direkte Beantwortung statt Ausgabe einer Textstelle
taq_3	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_4	Subparagraf	falscher Kontext
taq_5	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_6	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_7	Subparagraf	falscher Kontext
taq_16	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_4	Titel	Titel reicht nicht aus
taq_6	Titel	Titel reicht nicht aus
taq_8	Titel	Titel reicht nicht aus
taq_9	Titel	Titel reicht nicht aus
taq_10	Titel	Titel reicht nicht aus
taq_12	Titel	Titel reicht nicht aus
taq_13	Titel	Titel reicht nicht aus
taq_14	Titel	Titel reicht nicht aus
taq_15	Titel	Titel reicht nicht aus
taq_16	Titel	Titel reicht nicht aus

Tabelle 29: *Vermutete Fehlerursachen in der Verwendung von Mistral Large im Versuch 2. Betratet werden nur die TAQs, bei denen die Antwort von Mistral Large in den zwei Durchläufen übereinstimmten.*

taq_id	**Textebene**	**Fehlerursache**
taq_2	Paragraf	Falscher Kontext
taq_6	Paragraf	Antwort nur implizit enthalten
taq_7	Paragraf	Falscher Kontext
taq_16	Paragraf	Direkte Beantwortung statt Ausgabe einer Textstelle
taq_1	Subparagraf	Falscher Kontext
taq_5	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_9	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_16	Subparagraf	Zurückgegebene Textstelle trifft Thema sehr nah, aber falsch
taq_4	Titel	Titel reicht nicht aus
taq_6	Titel	Titel reicht nicht aus
taq_9	Titel	Titel reicht nicht aus
taq_10	Titel	Titel reicht nicht aus
taq_11	Titel	Themengruppe verzerrt Antwort
taq_12	Titel	Titel reicht nicht aus
taq_13	Titel	Titel reicht nicht aus
taq_14	Titel	Titel reicht nicht aus
taq_15	Titel	Titel reicht nicht aus
taq_16	Titel	Titel reicht nicht aus

Tabelle 30: *Titel, zu denen Mistral Large in den verschiedenen Durchläufen die relevante Textstelle, die irrelevante Textstelle oder beides zurückgegeben hat.*

Titel	**Relevante Textstelle zurückgegeben**	**Irrelevante Textstelle zurückgegeben**
Module und Modulgruppen des Masterstudiengangs Computing in the Humanities	×	
Anerkennung von Studienzeiten, Prüfungsleistungen und Praktikumsleistungen	×	×
Studiendauer und Studienumfang	×	×
Bestehen von Modulen und Wiederholung von Modulprüfungen	×	
Form und Durchführung von Prüfungen	×	
Form und Bewertung der Bachelor- oder Masterarbeit		×
Zulassung zur Bachelor- oder Masterarbeit		×
Themengebiete für die Masterarbeit im Masterstudiengang Computing in the Humanities		×
Modulhandbuch		×