Universität Regensburg

Analyse der Informationsbedürfnisse in Conversational Search im Bereich DIY

Bachelorarbeit im Fach Medieninformatik am
Institut für Information und Medien, Sprache und Kultur (I:IMSK)

Vorgelegt von: Elena Frank
Adresse: Hauptstraße 23, 95671 Bärnau
E-Mail (Universität): elena.frank@stud.uni-regensburg.de
E-Mail (privat): frank_elena15@web.de
Matrikelnummer: 2113049
Erstgutachter: PD Dr. David Elsweiler
Zweitgutachter: Prof. Dr. Niels Henze
Betreuer: Alexander Frummet M.Sc.
Laufendes Semester: SS 2024
Abgegeben am: 26.08.2024

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Diese Bachelorarbeit untersucht die Generalisierbarkeit der Taxonomie von Frummet et al. (2022) aus dem Bereich Kochen auf den Bereich Do It Yourself (DIY). Basierend auf dieser bestehenden Taxonomie wird eine neue für den Bereich DIY entwickelt. Die zentrale Forschungsfrage lautet: Wie gut lässt sich die Taxonomie von Frummet et al. (2022) auf andere prozedurale Bereiche anwenden?“. Um diese Frage zu beantworten, wurde eine geringe Menge an manuell annotierten Daten genutzt, um mit Hilfe des Few-Shot Promptings mit einem Large Language Model die Informationsbedürfnisse im DIY-Bereich vorherzusagen und eine größere Menge an Annotationsdaten zu erzeugen. Die Ergebnisse zeigen, dass die Koch-Taxonomie erfolgreich auf den DIY-Bereich übertragen werden konnte, was die Generalisierbarkeit bestätigt. Während die Gesamtleistung von GPT-4 mit einem F1-Score von 93 % überzeugte, zeigten sich bei der Vorhersage einiger Kategorien jedoch Herausforderungen. Der Vergleich der beiden Taxonomien zeigte signifikante Unterschiede in den Verteilungen der Informationsbedürfnisse, was darauf hinweist, dass in den beiden Domänen unterschiedliche Schwerpunkte bezüglich der Informationsbedürfnisse liegen. Diese Arbeit legt damit den Grundstein für die flexible Anpassung der Koch-Taxonomie auf andere prozedurale Bereiche und bietet Ansatzpunkte für zukünftige Forschungen, die die Generalisierbarkeit weiter prüfen und spezifizieren könnten.

Abstract

This bachelor thesis examines the generalizability of Frummet et al.'s (2022) taxonomy from the cooking domain to the Do It Yourself (DIY) domain. Based on this existing taxonomy, a new one is developed for the DIY domain. The central research question is: "How well can the taxonomy of Frummet et al. (2022) be applied to other procedural domains?". To answer this question, a small amount of manually annotated data was used to predict the information needs in the DIY domain using Few-Shot Prompting with a Large Language Model and to generate a larger amount of annotation data. The results show that the Cook-taxonomy could be successfully transferred to the DIY domain, confirming its generalizability. However, while the overall performance of GPT-4 was convincing with an F1-score of 93%, the prediction of some categories showed challenges. The comparison of the two taxonomies showed significant differences in the distributions of information needs, indicating that the two domains have different priorities in terms of information needs. This work thus lays the foundation for the flexible adaptation of the taxonomy from the cooking domain to other procedural domains and offers starting points for future research that could further test and specify its generalizability.

1 Einleitung

ChatBots, VoiceBots, oder digitale Sprachassistenten - mittlerweile gibt es viele verschiedene Formen der Conversational Artificial Intelligence (AI), welche eine intuitive und natürliche Mensch-Maschine-Interaktion, besser gesagt Mensch-Maschine-Kommunikation ermöglichen und im alltäglichen Leben integriert sind. Diese Conversational Assistants können anhand einer dialogbasierten Suche – auch Conversational Search genannt - in den verschiedensten Bereichen weiterhelfen und erleichtern somit den Alltag. Sprachassistenten können beispielsweise Einkaufsliste schreiben, Musik wiedergeben oder Fragen zum Wetter beantworten. Auf was solche Systeme allerdings noch nicht spezialisiert sind, sind Anleitungen, die während eines Prozesses unterstützen, wie beispielsweise während des Kochens ein Rezept anzuleiten oder auch bei Do It Yourself (DIY) Projekten wie ein Regal selbst du bauen oder einen Garten anzulegen. Damit solche Systeme die Anfragen und somit Informationsbedürfnisse der Nutzer in einem solchen Kontext verstehen, ist es unerlässlich diese zu untersuchen und zu analysieren.

Eine präzise Textklassifikation spielt im Bereich der Conversational AI eine bedeutende Rolle, da sie dazu beiträgt die Anfragen und Informationsbedürfnisse der Nutzer korrekt zu erfassen, zu verarbeiten und richtig darauf zu reagieren. Damit Sprachmodelle für spezifische Aufgaben wie die Textklassifikation eingesetzt werden können, sind in der Regel große Datenmengen für das Training und Finetuning dieser erforderlich (Wen et al., 2017). Die notwendige Datenannotation ist somit sehr kosten- und zeitintensiv. Large Language Models wie GPT-3 (Brown et al., 2020) oder GPT-4 (OpenAI et al., 2024) sollen durch ihr Kontextverständnis und hohe Leistungsfähigkeit bei diesem Problem helfen (Tan et al., 2024).

Das Ziel dieser Bachelorarbeit ist die Generalisierbarkeit der Taxonomie von Frummet et al. (2022) aus dem Bereich Kochen zu untersuchen. Demnach dient diese als Vorlage, um eine neue Taxonomie für den Bereich DIY zu erstellen. Diese wird mit einer geringen Menge manuell annotierter Daten genutzt, um mit

Hilfe des Few-Shot Promptings und einem Large Language Model die Informationsbedürfnisse in dem Bereich DIY vorherzusagen und damit eine große Menge an Annotationsdaten zu erzeugen. Die zentrale Forschungsfrage lautet: „Wie gut lässt sich die Taxonomie von Frummet et al. (2022) auf andere prozedurale Bereiche anwenden?“. Es wird untersucht inwiefern eine Taxonomie aus einem anderen Bereich nach Anpassung an den Kontext zur Vorhersage der Informationsbedürfnisse genutzt werden kann und ob die Ergebnisse zufriedenstellend sind, sodass mit einem Large Language Model eine große und relevante Datenmenge erzeugt werden kann. Dies würde die Textklassifikation in der Conversational Search verbessern und den Prozess der Entwicklung von Conversational Assistants erheblich beschleunigen.

Im nächsten Kapitel wird zunächst ein Überblick über die grundlegenden Themen dieser Arbeit gegeben. In Kapitel 3 wird die Methodik beschrieben durch welche der Aufwand der Datenannotation reduziert werden soll. Die Ergebnisse und Diskussion dieser folgen in Kapitel 4 und 5. Anschließend werden die Limitationen dieser Arbeit aufgezeigt, sowie eine Schlussfolgerung sowie Ideen für zukünftige Arbeiten erwähnt.

2 Stand der Technik und verwandte Arbeiten

Dieses Kapitel gibt einen Überblick über die grundlegenden Themen, auf denen diese Bachelorarbeit basiert. Zunächst folgt on 2.1 eine kurze Einführung in die Conversational AI sowie ihrer mittlerweile breiten Anwendung. In 2.2 werden die Herausforderungen der Analyse der Informationssuche und demnach Informationsbedürfnisse erläutert sowie Studien diesbezüglich im Kontext der Conversational Search genannt. Die Textklassifikation, welche in dieser Arbeit Anwendung findet, wird in 2.3 beschrieben. Auch die Herausforderungen dieser Methodik des maschinellen Lernens im Kontext der Conversational Search werden genannt sowie State-of-the-Art Sprachmodelle, die diese Probleme lösen können. Eine Strategie, die sich für solche Situationen bewährt hat, wird anschließend in 2.4 erläutert.

2.1 Conversational AI

Die Conversational AI ist eine sich kontinuierlich weiterentwickelnde Technologie, welche die Mensch-Maschine-Interaktion in den letzten Jahrzenten deutlich verändert hat, da sie eine Kommunikation zwischen Mensch und Maschine in natürlicher Sprache ermöglicht. Die Grundlage, um eine Unterhaltung auf diese Art und Weise zu ermöglichen, stellt Natural Language Processing (NLP) dar, was die natürliche Sprachverarbeitung meint. Zusammen mit Natural Language Understanding (NLU) – dem Verstehen natürlicher Sprache – und Natural Language Generation (NLG) – dem Generieren natürlicher Sprache – wird es Systemen ermöglicht Unterhaltungen in natürlicher Sprache zu führen.

Bereits in den 1960ern begann die Entwicklung dieser Technologie: Der Chatbot ELIZA (Weizenbaum, 1966) gilt als erstes Beispiel eines Computerprogramms, welches eine Unterhaltung mit Menschen in natürlicher Sprache führen kann. Der Nutzer tippt eine oder mehrere Aussagen in natürlicher Sprache ein, worauf diese Eingabe von ELIZA auf Schlüsselwörter überprüft wird, um dadurch mit Hilfe darauf basierenden Regeln eine Antwort zu generieren. ELIZA ahmt einen Psychotherapeuten nach und erweckt bei den Nutzern den Eindruck ihn verstehen zu können, was dem Programm menschliche Eigenschaften gibt.

Mittlerweile gibt es eine Vielzahl unterschiedlicher Systeme in der Conversational AI, welche auch Conversational Agent oder Conversational Assistant genannt werden. Diese verarbeiten die natürliche Sprache sowohl text- als auch sprachbasiert. Textbasierte Systeme verarbeiten die natürliche Sprache in schriftlicher Form und umfassen Chatbots wie ChatGPT[1], bei welchen der Nutzer durch Textnachrichten mit dem System kommuniziert und auch die Antwort in schriftlicher Form erhält, da die Konversation als Chat aufgebaut ist.

Sprachbasierte Systeme verstehen und verarbeiten die gesprochene Sprache des Menschen und generieren darauf passende Antworten. Bekannte Beispiele

[1] https://chatgpt.com/

sind unter anderem die digitalen Sprachassistenten Amazon Alexa[2], Siri[3] oder Google Assistant[4].

Die Conversational Assistants sind mittlerweile zunehmend in Websites integriert, um dadurch beispielsweise den Kundenservice zu unterstützen. Die Anfragen der Nutzer können dadurch schnell und vor allem rund um die Uhr beantwortet werden und somit die Benutzererfahrung verbessern. Beim Online-Shopping sollen die Systeme den Einkauf der Nutzer unterstützen und positiv beeinflussen (Balakrishnan & Dwivedi, 2024; Cui et al., 2017). In vielen Branchen wird die Technologie aktuell in der Forschung behandelt, wie beispielsweise in der Industrie (Saka et al., 2023), aber auch in sensibleren Bereichen wie der Medizin (Tu et al., 2024; Zhang & Boulos, 2023) oder des Online-Bankings (Kim & Song, 2024).

Doch vor allem auch für den privaten Gebrauch gewinnen die Conversational Assistants immer mehr an Beliebtheit, da sie in den verschiedensten Bereichen im Alltag unterstützen können. Obwohl sie bereits in vielerlei Hinsichten helfen können, sind sie auf einen Bereich noch nicht spezialisiert: den prozeduralen Bereich. Eine Anleitung während des Kochens zu führen oder durch den Prozess eines DIY-Projekts wie Gartenprojekte gehört noch nicht zu den Fähigkeiten dieser intelligenten Systeme. Damit diese in diesem Kontext funktionieren und die Anfragen und demnach Informationsbedürfnisse der Nutzer verstehen und verarbeiten können, ist es unverzichtbar diese zu untersuchen und zu analysieren.

2.2 Informationsbedürfnisse in der Conversational Search

Um eine Informationslücke zu füllen und das dadurch entstehende Informationsbedürfnis zu stillen, wird zusätzliches Wissen benötigt, um somit ein Problem lösen, eine Aufgabe erledigen oder eine Frage beantworten zu können. Dieses Ziel beabsichtigt der Prozess des Information Seekings, nämlich Informationen

[2] https://developer.amazon.com/de-DE/alexa
[3] https://www.apple.com/siri/
[4] https://assistant.google.com/

zu suchen, um ein Informationsbedürfnis zu befriedigen (Case, 2007). Eine offizielle Definition für den Begriff Informationsbedürfnis gibt es nicht, was auch Forsythe et al. (1992) beschreibt:

> no explicit consensus exists in the literature regarding the meaning of the central concept of "information need." … In effect, "information need" has been defined according to the particular interests and expertise of various authors. (Forsythe et al., 1992, S. 182)

Schon lange beschäftigen sich Wissenschaftler mit den Informationsbedürfnissen des Menschen und entwickeln Methoden, um diese zu analysieren und zu verstehen (Devadason & Lingam, 1997; Taylor, 1962; Wilson, 1981). Auch die Art und Weise wie Informationen gesucht werden, wurde schon in vielen Studien untersucht und analysiert (Kamvar & Baluja, 2006). Die Informationssuche veränderte sich durch die *Web Search* mit Suchmaschinen, da sich das Verhalten der Nutzer wesentlich von den Annahmen unterschiedet, die in der Literatur beschrieben werden (Silverstein et al., 1999, S. 6). Auch durch Smartphones und die dadurch entstehende *Mobile Search* hat sich die Suche nach Informationen verändert (Song et al., 2013), da somit die Möglichkeit besteht in unerwarteten Situationen (Church & Oliver, 2011, S. 75) nach Informationen zu suchen wie etwa einem Restaurant oder einer Bar in der Nähe. Church und Oliver (2011) nennen beispielsweise Zeit und Ort als starke Faktoren, welche das Informationsbedürfnis beeinflussen. Die Fähigkeit von Systemen die natürliche Sprache nicht nur in Textform, sondern auch in gesprochener Form zu verstehen, veränderte das Verhalten der Informationssuche der Nutzer (Guy, 2016; Yi & Maghoul, 2011).

Mit der Analyse direkt in der von Conversational Search beschäftigten sich bereits viele Studien. Bunt et al. (2017) erstellten eine domänenunabhängige Taxonomie, anhand derer in einem Dialog entstehende Äußerungen in verschiedene Kategorien geordnet werden können. Frummet et al. (2022) untersuchten die Informationsbedürfnisse während des Kochens und analysierten diese detailliert, um dadurch eine hierarchische Taxonomie zu erstellen. Auch Choi et al. (2022) fokussierten sich auf den Kontext Kochen sowie DIY-House Improvement und sammelten auf diese Bereiche bezogene Daten mit Hilfe einer Wizard-of-Oz-Studie, bei der jeder Teilnehmer einer von zwei Gruppen zugeteilt war. Die

dadurch erhobenen Dialoge bestehen aus Konversationen zwischen einem *Student*, der eine spezifische Aufgabe auszuführen hatte und einem *Teacher*, der über das erforderliche Fachwissen verfügte und Ersteren während des Prozesses anleitete.

Choi et al. (2022) und auch die Alexa Prize TaskBot Challenge 2021[5], bei welcher verschiedene Teams Conversational Assistants für die Bereiche Kochen und DIY entwickelten, verdeutlichen die Verknüpfung dieser beiden prozeduralen Bereiche, worauf auch Frummet et. al (2022, S. 25) hindeuten. Doch bisher gibt es keine detaillierte Taxonomie für DIY, welche vergleichbar mit der von Frummet et al. (2022) wäre.

2.3 Textklassifikation

Um die Informationsbedürfnisse in einem bestimmten Bereich untersuchen und analysieren zu können, müssen die relevanten Daten mit Hilfe der Textklassifikation in Kategorien eingeteilt werden. Diese Kategorisierung dient als Grundlage für die Erstellung einer Taxonomie. Die Textklassifikation ist eine typische Anwendung in Bereich NLP, bei der Daten von einem Sprachmodell klassifiziert werden (Li et al., 2021; Young et al., 2019). Eine präzise Textklassifikation ist im Kontext der Conversational AI besonders bedeutsam, um damit die Informationsbedürfnisse und damit Anfragen der Nutzer richtig zu verstehen, zu verarbeiten und entsprechend darauf reagieren zu können. Um Sprachmodelle für solche bestimmte Aufgaben wie Textklassifikation anwenden zu können, erfordert es für das Training und Finetuning dafür meist sehr große Datenmengen (Wen et al., 2017), was oftmals einen aufwändigen Prozess darstellt.

Durch die genannten Studien wird deutlich, dass nicht nur ein tiefgreifendes Verständnis der Informationsanforderungen, sondern vor allem große Datenmengen nötig sind, damit Systeme auf die speziellen Anfragen beziehungsweise Informationsbedürfnisse der Nutzer trainiert und spezialisiert und somit Conversational Assistants für diesen Bereich entwickelt werden können.

[5] https://www.amazon.science/academic-engagements/ten-university-teams-selected-to-participate-in-alexaprize-taskbot-challenge

2.3.1 Transformer-Modelle

Ein wesentlicher Fortschritt in NLP ist die Transformer-Architektur (Vaswani et al., 2017), auf welche mittlerweile viele bekannte Sprachmodelle basieren. Diese Architektur ermöglichte einen Fortschritt bezüglich der Fähigkeiten von Maschinen die natürliche Sprache zu verstehen und zu verarbeiten, da diese komplexe Beziehungen in Daten erkennen, wie auch das Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Das Sprachmodell wurde auf einem riesen Textkorpus vortrainiert und umfasst in seinem Basismodell insgesamt 110 Millionen Parameter. Es zeichnet sich durch sein bidirektionales Kontextverständnis aus: es ist also dazu in der Lage sowohl links-, als auch rechtsgerichtet den Kontext zu berücksichtigen, sodass die Bedeutung eines Wortes sowohl durch die vorangehenden, als auch durch die nachfolgenden Wörter präziser erfasst werden kann. Durch Feinabstimmung (Finetuning) kann BERT in verschiedenen Anwendungsbereichen eingesetzt werden wie beispielsweise schon erwähnt in der Textklassifikation. Weitere darauf aufbauende Sprachmodelle wie RoBERTa (Robustly Optimized BERT Pretraining Approach) erweiterte das Grundlagenmodell durch den Einsatz größerer Datenmengen und längerer Trainingseinheiten (Y. Liu et al., 2019). ALBERT (A Lite BERT for Self-supervised Learning of Language Representations) optimierte die Modellarchitektur, um dadurch den Speicherverbrauch zu reduzieren und gleichzeitig die Geschwindigkeit des Trainings zu erhöhen (Lan et al., 2020). Doch solche Sprachmodelle wie BERT benötigen für die Anwendung der Textklassifikation große Datenmengen für das Training, was in realen Szenarien oftmals nicht der Fall ist.

2.3.2 GPT-Modelle

Während Large Pre-Trained Models wie BERT große Datenmengen benötigen, um auf eine Aufgabe spezialisiert zu werden, ist das bei Large Language Models anders. Generative Pre-Trained Transformer (GPT)-Modellen, die ebenso auf der Transformer-Architektur basieren (Radford et al., 2018), ist ein Beispiel dafür. Ein bedeutender Fortschritt dieser Modelle war im Jahr 2020, als GPT-3 (Generative Pre-trained Transformer 3) von OpenAI vorgestellt wurde (Brown et al., 2020). Das Sprachmodell besitzt 175 Milliarden Parameter und ist demnach eines der

größten und leistungsfähigsten Sprachmodelle. Die Revolution dieses Modells ist, dass es in der Lage ist sich ohne umfangreiches (erneutes Pre-)Training oder spezifisches Finetuning auf verschiedene Aufgaben zu spezialisieren, beispielsweise durch das Zero-Shot oder Few-Shot Prompting. Ein weiteres Merkmal, das die GPT-Modelle auszeichnet, ist deren Fähigkeit zur Textgenerierung in natürlicher Sprache. Die weiterentwickelte Version GPT-4 (OpenAI et al., 2024) schlägt GPT-3 in vielen verschiedenen Aufgabenbereiche wie „reasoning, knowledge retention, and coding" (OpenAI et al., 2024, S. 44). GPT-4 kann nicht nur Text, sondern sogar Bilder als Eingabe verarbeiten und Antworten in Form von Text bereitstellen, was die Weiterentwicklung des zuvor schon sehr leistungsstarken GPT-3-Modells deutlich zeigt. Diese beiden effizienten Sprachmodelle werden in Conversational Assistants integriert, um die Möglichkeit der Conversational Search zu bieten und somit natürliche Konversationen mit Nutzern zu führen, wodurch sie bedeutende Fortschritte in der Entwicklung beziehungsweise Weiterentwicklung in NLP und der künstlichen Intelligenz darstellen.

Die GPT-Modelle wurden bereits in vielen Studien erfolgreich eingesetzt, auch in der Datenannotation. Da bei NLP-Anwendungen, besonders im Bereich der Datenannotation, oftmals eine große Datenmenge nötig ist, wird auf Crowdworking-Plattformen oder Annotationsexperten zurückgegriffen, welche, abhängig von der erforderten Datenmenge, Kosten verursachen. Gilardi et al. (2023) hebt die Effizienz von ChatGPT in so einer Situation hervor, da dadurch enorme Kosten eingespart werden. Das GPT-Modell ist um ein Vielfaches günstiger als das Crowd-Sourcing und liefert zudem auch noch bessere Ergebnisse. Auch Ding et al. (2023) zeigen, dass GPT-3 an die Fähigkeiten des Menschen im Bereich Datenannotation herankommt, da in deren Studie das Sprachmodell mit einer Accuracy von knapp 88% um nicht einmal 8% niedriger war als die des Menschen. Kosten und Zeit, welche bei einer manuellen Datenannotation sehr hoch sein können, konnten ebenso durch das Modell eingespart werden, da es viel günstiger als die manuelle Annotation durch den Menschen ist.

Auf Grund der Eigenschaft dieser leistungsstarken Sprachmodelle, dass sie ohne umfangreiches Training oder Finetuning auf Aufgaben spezialisiert werden

können, eignen sich diese Modelle sehr gut, wenn keine große Datenmenge vorhanden ist oder der Prozess der Datensammlung enorme Kosten verursachen würde. Durch solche Sprachmodelle können - ohne der Notwendigkeit vieler Beispieldaten für das Training oder Finetuning - viele klassifizierte Daten gesammelt werden. Mit diesen können dann letztlich Klassifikatoren für den Bereich Conversational Search trainiert und spezialisiert werden, um damit Dialogsysteme entwickeln zu können.

2.4 Prompt Engineering und Prompting

Prompt Engineering ist ein Forschungsbereich in der Human-Computer-Interaction (HCI), der sich auf die Gestaltung von Prompts für LLMs bezieht und untersucht wie die Kommunikation zwischen Mensch und künstlicher Intelligenz optimiert werden kann (Oppenlaender, 2023). Prompting bedeutet einem LLM eine präzise Anleitung, bestenfalls mit Kontext, in natürlicher Sprache als Input einzugeben wie dessen Verhalten sein soll. Anders gesagt gibt diese Anweisung dem LLM vor wie es die Aufgabe bearbeiten soll, also was dessen Antwort beziehungsweise Output ist (Amatriain, 2024). Somit bietet diese Methode eine flexible und vor allem effiziente Anpassung eines Sprachmodells an bestimmte Aufgaben.

Je klarer und detaillierter die Beschreibung der zu erfüllenden Aufgabe und damit die Anweisung beziehungsweise Prompt, der dem Modell übergeben wird, desto besser erfüllt dieses die Anforderungen des Nutzers (Bsharat et al., 2024). Durch solche Anleitungen wird das Sprachmodell dazu angeleitet eine präzise Antwort zu generieren, ohne der Erfordernis großer Rechenleistungen für das Training des Sprachmodells.

Neben der Anwendung bei text-to-text-Aufgaben findet das Prompting auch bei text-to-image seine Anwendung. Das von OpenAI vorgestellte Modell DALL-E (Ramesh et al., 2021), welches ebenfalls auf der Transformer-Architektur basiert, kann durch Texteingaben Bilder generieren. Diese Fähigkeit verdeutlicht das Potenzial des Promptings, wodurch generative Sprachmodelle präzise und zielgerichtet gesteuert werden können. Eine Studie von Gao et al. (2021), in der die Leistungsfähigkeit des Promptings mit der des Finetunings verglichen

wurde, berichtet von besseren Ergebnissen durch das Prompting, was die Fähigkeiten und vor allem Leistungsfähigkeiten dieser Methodik nochmals verdeutlicht. Mosbach et al. (2023) heben als weiteren Vorteil hervor, dass Prompting im Gegensatz zum Finetuning keine speziellen Fachkenntnisse erfordert, um effektiv eingesetzt zu werden und zudem kein Training oder Finetuning benötigt, was sowohl zeit-, als auch kostenintensiv ist.

Large Language Models in Verbindung mit Prompting wurden bereits in verschiedenen Anwendungsbereichen von NLP angewandt wie beispielsweise der Sentimentanalyse (Bu et al., 2024; Y. Wang & Luo, 2023) oder Übersetzung (Yamada, 2023) und erwies sich als effiziente Methode. Auch in der Datenannotation, bei der die hohe Präzision und Geschwindigkeit von großer Bedeutung ist, erbrachte diese Strategie bisher sehr gute Leistungen (Gilardi et al., 2023; Huang et al., 2023; Toney-Wails et al., 2024).

Doch das Prompt Engineering bringt auch Herausforderungen mit sich, da einige Merkmale beachtet und in dem Prompt berücksichtigt werden müssen. Ein Nachteil dieser Methodik ist die Empfindlichkeit gegenüber den Prompts. Selbst kleine Veränderungen dessen können deutlich Einfluss auf die Leistung des Modells nehmen. Doch Gu et al. (2023) und auch Sun et al. (2023) fanden heraus, dass durch das In-Context-Learning, beziehungsweise Few-Shot Prompting, bei dem also Demonstrationsbeispiele übergeben werden, diese Empfindlichkeit reduziert werden kann.

2.5 Forschungsfragen

Durch die vorherigen Kapitel wird deutlich, dass für die Conversational Search eine tiefe Analyse der Informationsbedürfnisse, vor allem aber große Datenmengen notwendig sind, um solche intelligente Conversational Assistants zu entwickeln. Dieser Prozess kann sehr zeit- und auch kostenintensiv sein, weswegen State-of-the-Art Sprachmodelle, zusammen mit einer effizienten Strategie, dieses Problem lösen können.

Da Conversational Assistants in prozeduralen Bereichen noch nicht spezialisiert sind, was daran liegen könnte, dass dafür große Datenmengen notwendig sind, soll in dieser Arbeit eine Methodik angewandt werden, mit dessen Hilfe das

Problem der aufwändigen Datenannotation umgangen wird. Die in Abschnitt 2.2 genannten Studien verdeutlichen die Verknüpfung der beiden Domänen Kochen und DIY und deuten darauf hin, dass DIY mit dem Bereich Kochen zusammenhängend sein könnte. Deswegen wird untersucht ob die Taxonomie aus dem Bereich Kochen von Frummet et al. (2022) für andere prozedurale Bereiche generalisierbar ist, um eben dadurch den Aufwand der Datenannotation deutlich zu verringern. Dies wäre ein Fortschritt im Bereich der Textklassifikation in der Conversational Search, wodurch der Prozess in der Entwicklung von Conversational Assistants beschleunigt werden könnte. Es wird eine Methodik angewandt, welche kosten- und zeitsparend ist, und durch die mit Hilfe der Taxonomie-Vorlage und einem Large Language Model (LLM) Informationsbedürfnisse vorhergesagt und dadurch eine große Menge an Annotationsdaten gesammelt werden können.

Durch die Probleme der aufwändigen Datenannotation ergeben sich folgende Forschungsfragen, welche in dieser Bachelorarbeit untersucht werden:

- „Wie gut lässt sich die Taxonomie von Frummet et al. (2022) auf andere prozedurale Bereiche anwenden?“
- „Wie genau können die Informationsbedürfnisse im Bereich DIY mit Hilfe einer Annotationsvorlage und eines LLMs vorhergesagt werden?“
- „Ist es möglich mit Hilfe einer Annotationsvorlage und einem LLM eine große Menge an annotierten Daten zu schaffen?“

Die erste Forschungsfrage ist hierbei die zentrale, wodurch dich sich die anderen ergeben.

3 Methodik

In diesem Kapitel wird die Methodik erläutert, welche die Beantwortung der Forschungsfragen ermöglichen soll, welche auf Grund des Problems der aufwändigen Datenannotation bestehen. Zunächst werden in Abschnitt 3.1 die Daten, welche die Grundlage dieser Arbeit sind, näher beschrieben. Zusammen mit der Taxonomie-Vorlage wird eine Stichprobe davon manuell annotiert und dadurch eine neue Taxonomie für den Bereich DIY erstellt, was in 3.2 beschrieben wird. In

dem letzten Abschnitt wird das Vorgehen beschrieben, um GPT-Modelle zu nutzen, um mit möglichst geringem Aufwand Informationsbedürfnisse zu klassifizieren und eine große Menge an annotierten Daten zu schaffen. was durch verschiedene Prompting Strategien ermöglicht werden soll.

Im Folgenden werden die einzelnen Schritte beschrieben, um einen Überblick über den Forschungsprozess dieser Bachelorarbeit zu geben. Damit sollen die daraus resultierenden Ergebnisse nachvollziehbar übermittelt werden.

3.1 Datengrundlage der Studie

Wie in Abschnitt 2.3 bereits erwähnt, ist die Datensammlung sehr zeitintensiv und aufwändig. Um diesem Aufwand entgegenzuwirken, wurden keine neuen Daten gesammelt, sondern stattdessen ein bereits existierender Datensatz ausgewählt, um diesen für die Annotation und damit zusammenhängender Analyse der Informationsbedürfnisse herzunehmen. Doch um einen geeigneten Datensatz zu finden, muss dieser anhand bestimmter Kriterien ausgewählt und entsprechend vorbereitet werden. Dabei ist sicherzustellen, dass dieser keine unnötigen Informationen enthält, welche für die Annotation und das Prompting nicht relevant sind.

3.1.1 Datensatz Wizard-of-Tasks

Ein entscheidendes Merkmal für die Auswahl des Datensatzes ist die Domäne, in welcher die Daten gesammelt wurden. Der von Choi et al. (2022) erstellte Datensatz Wizard of Tasks[6], welcher als erster Datensatz bestehend aus Konversationen zwischen einem Conversational (Task) Assistant und einem Nutzer in den Domänen Kochen und DIY-House Improvement gilt, stellt demnach eine gute Wahl dar. Im Folgenden wird der Begriff DIY-House Improvement mit DIY gleichgesetzt.

Doch nicht nur die Domäne, welche in dieser Studie behandelt wird, war entscheidend für die Auswahl dieses Datensatzes, sondern vor allem die Methode der Datensammlung. Wizard of Tasks wurde mit Hilfe eines Wizard-of-Oz-Ex-

[6] https://registry.opendata.aws/wizard-of-tasks/

periments erstellt, welche eine Situation zwischen einem Nutzer und einem Conversational Assistant in natürlicher Sprache imitiert. Dafür wurde Studie durchgeführt, bei welcher während eines Gesprächs ein Teilnehmer einen Conversational Assistant und ein anderer einen Nutzer imitiert. Jeder Teilnehmer war nur einer Gruppe zugeteilt, um sicherzustellen, dass die Qualität nicht beeinflusst wird (Choi et al., 2022, S. 3516), demnach also entweder der Gruppe *Teacher* (entspricht dem Conversational Assistant) oder *Student* (entspricht dem Nutzer). Diese Methode imitiert also genau die Frage-Antwort-Situation in welcher ein Nutzer mit einem Conversational Assistant in natürlicher Sprache interagiert, um eine Aufgabe erfolgreich zu absolvieren.

Ein Vorteil der Methodik, mit welcher die Daten gesammelt wurden, liegt darin, dass eine schnelle Datensammlung möglich wird, da keine Wartezeit auf die Antwort eines anderen Teilnehmers anfällt. Zudem kann ein Teilnehmer zügig an mehreren Aufgaben beziehungsweise in diesem Fall Gespräche teilnehmen beziehungsweise diese bearbeiten, was die Methode auch effizient macht und somit Kosten reduziert werden (Wen et al., 2017). Der Nachteil dieser Methode liegt darin, dass manche Teilnehmer keine ernstgemeinten Antworten liefern. Dem wurde allerdings entgegengewirkt, indem die Antworten überprüft und notfalls ein Teilnehmer für das Weiterarbeiten temporär blockiert wurde.

Zudem weisen die gesammelten Gespräche eine außerordentlich hohe Qualität mit einer Relevanz von über 97% auf, basierend auf einer Bewertung der Teilnehmer hinsichtlich Relevanz und Nützlichkeit der vorherigen Äußerung (Choi et al., 2022, S. 3518).

3.1.2 Vorbereitung der Daten

Um einen Datensatz für die Verarbeitung nutzen zu können, muss dieser oftmals vorbereitet werden, damit unnötige Daten entfernt werden und die Daten im richtigen Format und in der richtigen Struktur vorliegen. Der Datensatz ist als json-Datei mit mehreren Einträgen gespeichert, wobei jeder einen Dialog zwischen einem Benutzer (*Student*) und einem Conversational Assistant (*Teacher*) im

DIY-Bereich darstellt. Die Daten wurden aufbereitet, um eine einheitliche Struktur zu gewährleisten und die Daten für die anschließende Annotation mit zugehöriger Analyse vorzubereiten.

Demnach wurden nur die für diese Arbeit relevanten Daten aus der json-Datei extrahiert. Irrelevante Informationen wie beispielsweise IDs der Teilnehmer wurden entfernt, um dadurch den Fokus auf die relevanten Dialoge zu legen. Zusätzlich wurden die Daten in ein Pandas DataFrame umgewandelt, um die Datenmanipulation und -analyse mit der Python-Bibliothek Pandas (McKinney, 2010) zu ermöglichen. Das im maschinellen Lernen typische Preprocessing der Daten wie beispielsweise Tokenisierung, Lemmatisierung oder das Entfernen von Stoppwörtern wird hier allerdings nicht benötigt, da LLMs wie GPT-3 oder GPT-4 die natürliche Sprache verstehen können.

Durch diese Vorbereitungsschritte wird sichergestellt, dass die Daten in einem geeigneten Format vorliegen und nur relevante Informationen enthalten sind, um die Annotation und Analyse der Daten effektiv durchführen zu können.

3.1.3 Stichprobe

Innerhalb der Wizarf-of-Tasks-Studie wurden für den DIY-Bereich insgesamt 277 Gesprächsverläufe bestehend aus 10.169 Äußerungen gesammelt. Die Konversationen enthielten im Gegensatz zu denen im Bereich Kochen circa 20% mehr Äußerungen, was darauf hindeutet, dass DIY-Aufgaben aufwändiger sind und detaillierte Informationen erfordern, damit das Ziel erreicht wird (Choi et al., 2022, S. 5). Dies verdeutlicht die Komplexität der Informationsbedürfnisse in dieser Domäne und die notwendige Analyse dieser, damit die Entwicklung von Conversational Assistants für diesen Anwendungsbereich ermöglicht wird.

Im dieser Studie wird allerdings nur eine Stichprobe des Datensatzes behandelt, da eine Methode getestet wird, mit Hilfe derer mit möglichst geringem Aufwand eine große Menge an annotierten Daten gesammelt werden soll, um dadurch Zeit und Kosten zu reduzieren. Bei einer Studie von Figueroa et al. (2012), bei der die Performance eines Klassifikationsmodells bei steigender Trainingsdatengröße untersucht wurde, wurden 100-200 annotierte Daten als ausrei-

chende Startgröße genannt, weshalb eine Menge von 150 als Startgröße genommen wurde. Da die Äußerungen zusammenhängend pro Konversation betrachtet wurden, um bei der Annotation den bisherigen Verlauf und damit den Kontext des Gesprächs berücksichtigen zu können, wurden letztlich 171 Äußerungen seitens *Student* ausgewählt. Da jedoch manche Labels in diesen ersten Konversationen zu wenig vertreten waren, wurden noch mehr Gesprächsverläufe berücksichtigt, woraus schließlich 227 Äußerungen von *Student* manuell annotiert wurden.

3.2 Manuelle Annotation mit Hilfe der Taxonomie-Vorlage

Das Ziel der manuellen Annotation ist eine relativ kleine, aber dennoch repräsentative Menge an annotierten Daten zu sammeln, welche letztlich an ein LLM übergeben werden, sodass durch dieses dann die restlichen Daten klassifiziert werden und somit eine große Datenmenge entsteht. Vor allem aber geht es darum die Informationsbedürfnisse in der Domäne DIY zu untersuchen und ein detailliertes Verständnis dieser zu erlangen, um somit eine Taxonomie erstellen zu können. Dies ist für die Entwicklung der Conversational Assistants für solche prozeduralen Bereiche wie DIY von enormer Wichtigkeit.

Um dem Aufwand der Erstellung einer neuen Taxonomie entgegenzuwirken, wird die Taxonomie von Frummet et al. (2022) als Vorlage genutzt. In deren Studie mit 45 Teilnehmern wurde der Prozess des Kochens durchlaufen, angefangen bei der Rezeptsuche, bis hin zur Zubereitung des ausgewählten Gerichts. Dies wurde mit Hilfe eines Conversational Assistants, welcher vom Experimentleiter verkörpert wurde, durchgeführt. Dabei war die Studienumgebung jeweils bei den Teilnehmern zuhause, wodurch eine reale Situation simuliert werden konnte, welche entsteht, wenn ein Nutzer zusammen mit einem solchen digitalen Sprachassistenten ein Gericht zubereitet. Dadurch konnten sehr realitätsnahe Daten gewonnen werden, da die Umgebung für die Teilnehmer gewohnt war und dadurch das Gefühl einer künstlich erzeugten Situation minimiert wird. Diese Daten wurden im Anschluss transkribiert und annotiert, wodurch eine detaillierte Taxonomie entstanden ist: durch die verschieden detaillierten Kategorien ergibt sich eine Hierarchie, welche insgesamt aus den sechs Stufen Level 0 bis

Level 5 besteht. Pro Level gibt es eine unterschiedliche Anzahl an Kategorien, denen die Äußerungen der Studie zugeteilt sind. Bei Level 0 gibt es die beiden Kategorien *Fact* und *Competence*, bei Level 1 gibt es hingegen bereits elf Kategorien *Amount*, *Ingredient*, P*reparation*, *Cooking Technique*, *Recipe*, *Time*, *Equipment*, *Knowledge*, *Meal*,*Temperature* und *Miscellaneous*. In den darauffolgenden Level sind noch detailliertere Kategorien aufgeführt, welche sich immer mehr an die spezifische Situation anpassen. Level 2 besteht bereits aus 93 Kategorien, wodurch erkennbar ist, dass die Taxonomie pro Stufe immer detaillierter wird und immer passender auf die Äußerungen und somit auch auf die Domäne zugeschnitten ist. Jede Äußerung der Teilnehmer der Studie wurde von Level 0 bis 2 codiert, die darauffolgenden Level nur optional, da diese sehr detailliert auf die genaue Äußerung zugeschnitten waren. Dadurch, dass die Kategorien pro Stufe immer detaillierter werden, sind diese auch sehr domänenspezifisch. Frummet et al. (2022, S. 25) vermuten, dass die Level 0 und 1 ihrer Taxonomie auch auf andere prozedurale Bereiche anwendbar sind, weswegen diese in dieser Arbeit als Vorlage für die Annotation dienen. Demnach wird geprüft ob die Kategorien der ersten beiden Level der Domäne Kochen auf DIY übertragbar sind und im Idealfall sogar direkt übernommen werden können.

Der Prozess der manuellen Annotation wurde mit der Einarbeitung in die Daten begonnen, um einen Überblick über die Äußerungen der Stichprobe zu gewinnen. Zudem wurden die Definitionen der Kategorien, sowie einige Äußerungen der Arbeit von Frummet et al. (2022) untersucht, wodurch eine erste Einschätzung bezüglich der Übertragbarkeit der Koch-Kategorien auf DIY entstanden und ein erster Entwurf der angepassten Taxonomie erstellt worden ist. Wenn die Kategorien zu domänenspezifisch sind, sind diese - wenn möglich - für den Kontext passend umformuliert oder andernfalls aussortiert worden.

Anfangs war eine Multilabel-Klassifikation geplant, sodass also einer Äußerung auch zwei oder mehr Label zugeteilt werden können. Da allerdings durch erste Versuche des Promptings das Sprachmodell keine guten Vorhersagen der Kategorien erwies, wurde sich für eine Muliklassen-Klassifikation entschieden, also pro Äußerung ein Label. Zudem haben die Teilnehmer der Wizard-of-Tasks

Studie für ihre jeweilige Äußerung nur eine Intention angegeben, was darauf hindeutet, dass sie auch nur ein Informationsbedürfnis mit der Äußerung äußerten.

Während des iterativen Prozesses der Analyse und Annotation der Äußerungen wurden die Taxonomie und zugehörige Definitionen weiter angepasst. Falls eine Unsicherheit bezüglich einer Zuordnung einer Kategorie vorhanden war, wurden bei manchen Äußerungen Notizen festgehalten. Außerdem wurden nur Äußerungen klassifiziert, die als Anfrage an das System gewertet werden können und somit ein Informationsbedürfnis darstellen. Demnach wurden Beispiele wie „Thank you for the instructions!“ oder „I guess that's what I'll do then! I appreciate you trying to help me here!“ nicht berücksichtigt, da hier seitens des Nutzers keine Antwort vom System erwartet wird, wodurch es nicht als Anfrage gilt. Insgesamt wurde die Annotation so lange iterativ durchgeführt bis ein zufriedenstellendes Ergebnis einer passenden Taxonomie mit zugehörigen Äußerungen erreicht wurde.

3.3 Datenannotation durch Prompting-Strategien

Abbildung 1: Prozess des Promptings (eigene Abbildung)

Um zu überprüfen ob die Größe der Stichprobe zusammen mit den manuellen Annotationen ausreichend und repräsentativ genug für den gesamten Datensatz ist, wird diese einem Sprachmodell übergeben, damit dieses auf Basis der übergebenen Daten eine große Menge an Annotationsdaten sammeln kann. Mit dieser Methodik wird also überprüft, ob mit wenig annotierten Daten der Prozess der Datenannotation mit Hilfe von Sprachmodellen unterstützt werden kann. Dafür wird ein LLM eingesetzt, welches für Textklassifikation geeignet ist und somit den Äußerungen die Labels zuordnet.

Hierfür werden mit den von OpenAI entwickelten Sprachmodellen GPT-3 (Brown et al., 2020) und GPT-4 (OpenAI et al., 2024) zwei verschiedene Methoden

des Promptings, dessen Prozess in Abbildung 1 zu sehen ist, getestet. GPT-3 bewährte sich bisher bei der Annotation von Daten und gilt darin als kostengünstig und zeitsparend (Ding et al., 2023; S. Wang et al., 2021). Zusätzlich wird das darauf aufbauende Modell GPT-4 getestet, da dieses die weiterentwickelte Version ist. Zudem wird auch auf der offiziellen Homepage von OpenAI darauf hingewiesen, dass das weiterentwickelte Modell bessere beziehungsweise effizientere Ergebnisse erzielt[7].

Die beiden für diese Arbeit geeigneten Methoden des Promptings sind *Zero-Shot* und *Few-Shot* Prompting, da sich diese auf einen Bereich spezialisieren: „new tasks without extensive training" (Sahoo et al., 2024, S. 2). Diese beiden Strategien werden getestet, um letztlich mit der besseren den gesamten Wizard-of-Tasks Datensatz klassifizieren zu lassen und somit eine große Menge an annotierten Daten zu erhalten, welche für die Entwicklung von Conversational Assistants unerlässlich ist.

3.3.1 Prompt Engineering

Wie bereits in Abschnitt 2.4 erwähnt, meint Prompt Engineering die Entwicklung und Gestaltung eines Prompts, der einem LLM als Art Anleitung übergeben wird. In diesem iterativen Prozess wird dieser ständig weiterentwickelt beziehungsweise verfeinert, um eine mögliche Verbesserung in den Ergebnissen zu erreichen. Im Folgenden werden Guidelines dargestellt, mit deren Hilfe der notwendige Prompt erstellt wurde. Anschließend wird die Anwendung dieser an dem entstandenen Prompt gezeigt.

3.3.1.1 Guidelines des Prompt Engineerings

Um zu einem effektiven Ergebnis dieses Prozesses zu gelangen, gibt es Guidelines oder bestimmte Ansätze, welche Empfehlungen für das Design und die Struktur des Prompts bereitstellen (Amatriain, 2024; Bsharat et al., 2024).

Neben der Erstellung des Äußeren des Prompts gibt es auch verschiedene Ansätze auf welche Art und Weise der Prompt erstellt wird. Eine gängige Methode ist es den Prompt manuell zu erstellen. Da die GPT-Modelle fähig sind die

[7] https://platform.openai.com/docs/models/gpt-3-5-turbo

natürliche Sprache zu verstehen, kann der Kontext der Aufgabe bei der manuellen Erstellung präzise beschrieben werden, wodurch bessere Antworten seitens des Modells erzielt werden können.

Um dem LLM einen zielgerichteten Prompt übergeben zu können, müssen detaillierte Informationen bereitgestellt werden, welche die zu absolvierende Aufgabe präzise darstellen, Für die dafür notwendige äußere Struktur und Aufbau des Prompts wird sich an dem CO-STAR-Framework orientiert, welches vom Data Science & AI Team von GovTech Singapore entwickelt wurde (GovTech Data Science & AI Division, 2023). Dieses Framework dient als Anleitung, um einen effektiven Prompt zu erstellen, indem durch einen strukturierten Aufbau alle notwendigen Informationen für das LLM bereitgestellt werden. CO-STAR stellt sich aus sechs Elementen zusammen:

- C – Context: Hintergrundinformationen zum Thema der zu absolvierenden Aufgabe helfen dem System ein besseres Verständnis für die Situation zu haben.
- O – Objective: Eine klare Definition der Aufgabe, welche das System ausführen muss, hilft diesem eine passende und relevante Antwort zu generieren.
- S – Sytle: Der Schreibstil der generierten Antwort wird passend zur Aufgabe definiert, damit das System sich noch mehr an die spezifische Aufgabe beziehungsweise das Ziel dieser anpassen kann.
- T –Tone: Auch der Tonfall kann spezifiziert werden, indem der emotionale Kontext einbezogen und beispielsweise durch Adjektive wie humorvoll oder freundlich dargestellt wird.
- A – Audience: Das Erwähnen der Zielgruppe kann ebenso Einfluss auf die Antwort nehmen und diese spezifizieren, da somit die Komplexität der Sprache angepasst wird.
- R – Response: Die Form und Länge der gewünschten Antwort.

Durch diese Elemente werden alle für das LLM notwendigen Informationen abgedeckt, um den Kontext der Aufgabe zu verstehen und somit relevante Antworten zu liefern.

Auch die inhaltlichen Formulierungen im Prompt sind entscheidend für die Leistung des LLMs. Bsharat et al. (2024) erstellten eine ausführliche Anleitung zum inhaltlichen Design des Prompts. Insgesamt führen sie 26 Prinzipien auf, welche beachtet werden sollten, um den Prompt noch effektiver zu gestalten. Diese werden beim Prozess des Prompt Engineering berücksichtigt, sofern sie passend für das Ziel der Aufgabe - der Datenannotation - sind.

3.3.1.2 Prompt Engineering in dieser Arbeit

Mit Hilfe des CO-STAR Frameworks, was bereits in Abschnitt 3.3.1.1 definiert wurde, wurde zuerst die Struktur des Prompts festgelegt. Für jede Komponente wurde ein Text bereitgestellt, durch die Details und der Kontext der Textklassifikationsaufgabe überliefert wurden.

Für „Context" wurde eine kurze Einführung in das Thema definiert, bei der beschrieben wurde was sich in den Daten befindet und welche Situation bei der Sammlung dieser bestand. Demnach wurde erwähnt, dass die Daten Konversationen zwischen einem Conversational Assistant und einem Nutzer darstellen, welche zusammen den Prozess zur Durchführung eines DIY-Projekts durchlaufen. Zudem wurde der Hintergrund der Klassifikation angegeben, welcher also erklärt, dass die Äußerungen mit Hilfe einer Taxonomie annotiert werden müssen, damit die Informationsbedürfnisse der Nutzer identifiziert werden können.

Bei dem Teil „Objective" wurde die Aufgabenstellung bereitgestellt. Das LLM wird dazu aufgefordert die gegeben Äußerungen der Teilnehmer, welche den Nutzer verkörperten, zu annotieren. Damit das Sprachmodell einen Überblick über die Kategorien hat, wurden diese zusammen mit einer Definition dessen aufgeführt. Bei Level 0 wurden hier die beiden Kategorien zusammen mit ihren Unterkategorien definiert, damit das LLM eine klarere Vorstellung der beiden Typen von Level 0 hat. Zudem wurde darauf hingewiesen, dass sowohl die vorherigen Äußerungen der Nutzer selbst, als auch die des Conversational Assistants mit berücksichtigt werden müssen, da dies einen Einfluss auf die Kategorie des Informationsbedürfnisses haben kann. Auch Frummet et al. (2022) führten in ihrer Arbeit auf, dass der Kontext und auch der Konversationsverlauf bei einer Klassifikationsaufgabe mit berücksichtigt werden sollte, damit der nötige

Kontext mit in die Entscheidung der Kategorisierung der Äußerungen einfließen kann.

Da die Antwort des LLMs nur die einzelne Kategorie sein soll, wurde bei „Style“ die Anweisung ergänzt, dass es nur eine der aufgeführten Kategorien als Antwort nutzen soll. Das Element „Tone“ wurde jedoch nicht hinzugefügt, da dies bei einer einzelnen Kategorie als Ouput nicht notwendig ist.

Bei „Audience“ wurde angegeben, dass die Ergebnisse beziehungsweise die entstehenden Kategorien zur Analyse durch Experten benötigt werden, da Bsharat et al. (2024) in ihrer ausführlichen Anleitung für das Prompt Engineering erwähnten, dass dies die Leistung des Sprachmodells positiv beeinflusst.

Am Ende des Prompts wurde durch das Element „Response“ nochmal darauf hingewiesen, dass für jede Äußerung nur der Begriff des Informationsbedürfnisses ausgegeben werden soll. Zusätzlich wurden dem Prompt bei der Few-Shot Methode die Beispiel-In- und Beispiel-Outputs angehängt.
Dadurch, dass das Prompt Engineering ein iterativer Prozess ist, um die Anweisung weiter zu optimieren, wurden die einzelnen Inhalte der CO-STAR-Elemente immer wieder verändert und spezifiziert. Dadurch konnte festgestellt werden, dass zu viel Kontext auch einen negativen Effekt haben kann. Nachdem der Prompt übergeben wurde, wurden die Ergebnisse zwischen den verschiedenen Kategorien analysiert, indem beispielsweise die „false positives“ beziehungsweise „false negatives“ analysiert wurden. Durch diese Analyse wurden Einblicke gewonnen, welche Kategorien noch zu ungenau definiert sind, sodass das Modell keine klare Grenze zwischen zwei oder mehreren Kategorien erkennt und es deswegen eine falsche Kategorie für eine Äußerung wählt. Anhand dieser Analyse wurden die Definitionen der Kategorien in dem Prompt nochmals überarbeitet und neu beziehungsweise detaillierter definiert, damit die Definitionen klarer sind. Aufgrund der Prinzipien von Bsharat et al. (2024) wurde außerdem in „Objective“ ein Satz ergänzt, der dem LLM eine klare Rolle eines Experten zuweist, wodurch sich die Klassifikation verbessert hat. Laut Amatriain (2024) ver-

arbeiten Sprachmodelle direkte und entschlossene Formulierungen besser als zurückhaltende und freundliche, weshalb auch Ausrufezeichen im Prompt verwendet wurden, was tatsächlich die Leistung des Modells verbesserte.

Durch diese übersichtliche Struktur zusammen mit der präzisen Ausarbeitung des Inhalts werden dem LLM viele Details der Aufgabe bereitgestellt, sodass es sich dieser optimal anpassen kann, da der Kontext dazu bekannt ist.

3.3.2 Zero-Shot-Prompting

Bei dieser Strategie wird einem LLM nur der Prompt als Anleitung gegeben, in dem die Aufgabe und gewünschte Antwort definiert sind. Dem Sprachmodell werden jedoch keine Beispieldaten übergeben, von welchen es lernen könnte. Diese Strategie benötigt einerseits ein Modell mit einem hohen Sprachverständnis, andererseits vor allem einen präzisen und zielgerichteten Prompt, damit das Modell die erwartete Antwort ausgibt. Obwohl es naheliegend scheint, dass ein *Zero-Shot* Ansatz schlechter als ein *Few-Shot* performt, gibt es auch Fälle, bei denen die Methode ohne Übergabe von Beispielen sogar bessere Ergebnisse erzielen kann (Reynolds & McDonell, 2021). Die Vorteile dieser Strategie sind natürlich der geringere Aufwand an Zeit und Kosten im Gegensatz zur *Few-Shot* Methode, da nur der Prompt erstellt werden muss und keine Beispiele gesucht werden müssen.

3.3.3 Few-Shot-Prompting

Beim *Few-Shot* Prompting werden dem LLM Beispiele übergeben, welche sowohl den Input, als auch den gewünschten Output der von diesem zu absolvierende Aufgabe beinhalten. Dadurch kann das Sprachmodell die Vorgehensweise und somit die Gedanken des Menschen besser nachvollziehen welche Antwort genau erwartet wird. Doch welche Beispiele genau verwendet werden, kann einen Einfluss auf die Leistung des Modells haben, da dieses sehr sensitiv darauf reagiert (J. Liu et al., 2021; Lu et al., 2022). In Studien wurden bereits viele Einflussfaktoren untersucht, wie zum Beispiel die Anzahl und auch Anordnung der *Few-Shot* Beispiele, da dies einen Einfluss auf die Leistung des Modells haben kann (Lu et al.,

2022; Zhao et al., 2021). Eine zu große Anzahl an Beispielen kann einen negativen Einfluss auf das Sprachmodell haben (Pecher et al., 2024).

Deshalb werden die beiden Faktoren Anzahl und Anordnung auch in dieser Studie mit einbezogen, um herauszufinden durch welche Kombination derer die besten Ergebnisse erzielt werden können. Hierfür wurden die Anordnungen der Beispiele zusammen pro Kategorie verändert. Der Unterschied zwischen den verschiedenen Reihenfolgen der Kategorien machte nahezu keinen Unterschied aus (höchstens +0.004 in *Precision*, *Recall* und *F1-Score*). Die besten Ergebnisse bezüglich der Anzahl der Beispiele wurden bei der Vorhersage der Level 0 Kategorien mit $k = 20$, bei Level 1 mit $k = 4$ erreicht.

4 Ergebnisse

In den folgenden Abschnitten werden die Ergebnisse dargelegt, durch welche die in Abschnitt 2.5 aufgeführten Forschungsfragen behandelt werden.

4.1 Manuelle Annotation

Im Folgenden werden die Ergebnisse der manuellen Annotation aufgeführt. Wie bereits in 3.1.3 erwähnt, wurden dafür insgesamt 227 Äußerungen von *Student* kategorisiert. Durch den iterativen Prozess der Annotation wurde mit Hilfe der Vorlage aus der Koch-Domäne eine neue Taxonomie erstellt.

4.1.1 Taxonomie für die Domäne DIY

Level 0	**Fact**	**Competence**
Level 1	Equipment	Preparation
	Material	Manual Technique
	Miscellaneous	
	Instruction	
	Knowledge	
	Time	
	Amount	

Abbildung 2: Struktur der Taxonomie für den Bereich DIY

Die Taxonomie besitzt die Struktur eines Baumes, bestehend aus der oberen Ordnung Level 0 und der unteren Ordnung Level 1. Demnach stammt also jede Level 1 Kategorie von einer der beiden in Level 0, wodurch jedes Informationsbedürfnis sowohl einem Level 0, als auch einem Level 1 zugeordnet ist.

In Abbildung 2 ist eine Übersicht der Taxonomie mit den eher oberflächlichen Kategorien des Levels 0, und den zugehörigen detaillierteren des Levels 1 abgebildet. Die entstandene Taxonomie für die Domäne DIY besteht aus den beiden Level 0 Typen *Fact* und *Competence*, sowie den neun Level 1 Informationsbedürfniskategorien *Amount*, *Material*, *Preparation*, *Manual Technique*, *Instruction*, *Time*, *Equipment*, *Knowledge* und *Miscellaneous*. Während des Prozesses eines DIY-Projekts werden mögliche Projekte oder allgemeine Informationen zu DIY-Projekten erfragt, weshalb die Kategorie *Instruction* in dieser Domäne passend ist. Zudem werden Informationen zu Arbeitsmitteln und zugehörige Fakten wie Mengen oder Zeitangaben während des Prozesses benötigt, weshalb die Informationsbedürfnisse *Material* sowie *Amount*, *Equipment* und *Time* in der Taxonomie aufgenommen wurde. Fragen nach den einzelnen Schritten tauchen in jedem prozeduralen Bereich auf, weshalb *Preparation* als Informationsbedürfnis bezüglich der einzelnen Schritte passend ist, genauso wie Fragen bezüglich Definitionen oder Hintergrundwissen, wodurch die Kategorie *Knowledge* in der Taxonomie Verwendung findet. Da der Bereich DIY sehr spezifisch ist, können Situationen auftreten, in denen bestimmte Techniken oder Ausführungen nötig sind, die ein gewisses Vorwissen benötigen. Für solche Situationen wurde die Kategorie *Manual Technique* in die neue Taxonomie aufgenommen. Zuletzt wurde die Kategorie *Miscellaneous* hinzugefügt für Äußerungen, welche den anderen Kategorien nicht zugeordnet werden konnten.

4.1.2 Neuer Datensatz allgemein

Wie bereits in Abschnitt 3.1.3 erwähnt, beinhaltet der Datensatz Wizard-of-Tasks insgesamt 10.169 Äußerungen im DIY-Kontext, wovon 456 der Stichprobe angehören und der Anteil von *Student*, nämlich 227 (etwa 50%) Äußerungen, manuell

annotiert wurden. Jede Äußerung, welche aus einem oder mehreren Sätzen besteht, wurde auf Grund der Multiklassen-Textklassifikation nur einem Informationsbedürfnis zugeteilt.

Durchschnittlich sind einer Level 0 Kategorie 113.5 Äußerungen zugeteilt (*min* = 82.0, $x_{.25}$ = 97.75, $x_{.50}$ = 113.5, $x_{.75}$ = 129.25, *max* = 145, *sd* = 44.55). Bei Level 1 liegt der Durchschnittswert bei 25.22 Äußerungen pro Kategorie (*min* = 4.0, $x_{.25}$ = 5.0, $\tilde{x}_{.50}$ = 12.0, $x_{.75}$ = 16.0, *max* = 138, *sd* = 42.91). Eine Äußerung von *Student* hatte eine durchschnittliche Länge von 12.88 Wörtern (*min* =3.0, *max* = 54.0).

4.1.3 Intra- und Inter-Rater Reliabilität

Informationsbedürfnis	**Cohen's Kappa**
Equipment	0.86
Preparation	0.84
Material	0.52
Miscellaneous	0.39
Instruction	0.26
Knowledge	0.58
Manual Technique	0.22
Time	0.44
Amount	0.80

Tabelle 1: Cohen's Kappa Werte für jedes Informationsbedürfnis in Level 1

Dadurch, dass es Einflussfaktoren auf die Annotation einer Person gibt und diese eine subjektive Einschätzung ist, wurde zusätzlich ein zweiter Annotator hinzugezogen. Auch dieser hatte Einblick in die kompletten Konversationen, um den Kontext der Äußerungen zu verstehen. Der Ablauf der Annotation begann ebenso mit dem Einarbeiten in die Daten, um einen Überblick über die Äußerungen der Stichprobe zu gewinnen. Daraufhin wurden die 227 Äußerungen iterativ annotiert bis ein zufriedenstellendes Ergebnis erreicht wurde. Durch den Einsatz von zwei verschiedenen Annotatoren wird eine Vertrauenswürdigkeit untersucht bezüglich der Qualität der Annotation. Damit kann die Inter-Rater Reliabilität gemessen werden, also inwiefern die Annotationen zwischen den beiden

verschiedenen Annotatoren zufällig sind oder nicht. Auch hierfür wurde mit Hilfe Cohen's Kappa gemessen, die Ergebnisse sind in Tabelle 1 zu sehen. Gesamt bedeutet dies ein Cohen's Kappa Wert von 0.55, was auf ein moderates Ergebnis hindeutet (Landis & Koch, 1977).

4.1.4 Beschreibung der Info Needs

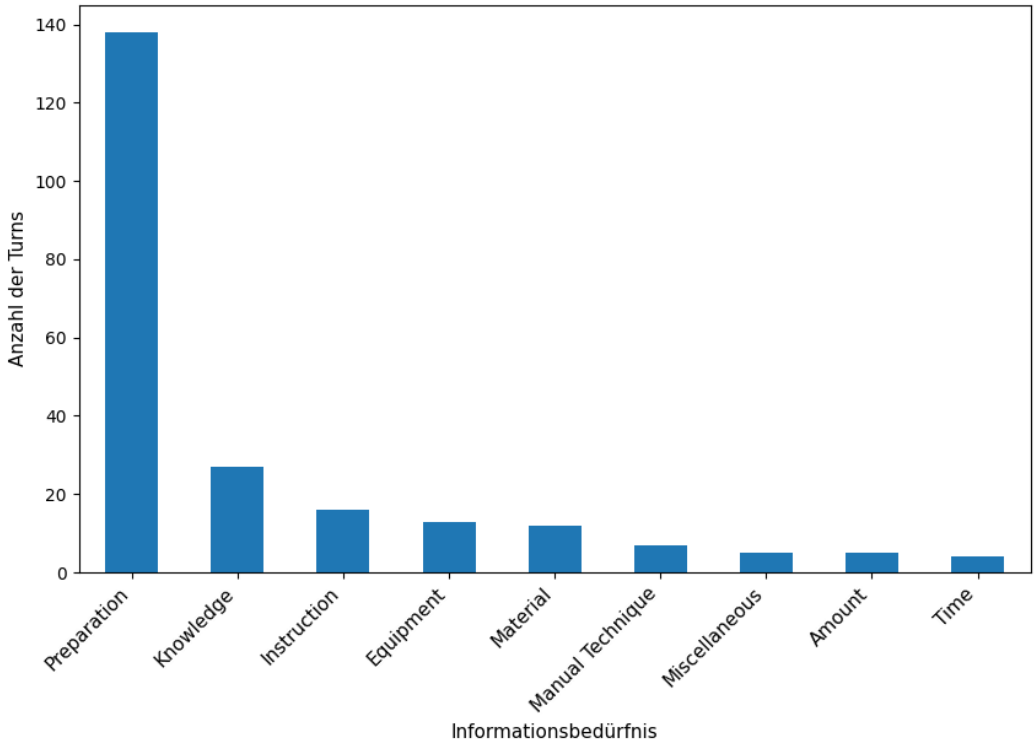

Abbildung 3: Verteilungen der Informationsbedürfnisse in der Stichprobe

Wie bereits in Abschnitt 4.1.1 aufgeführt, beginnt die Taxonomie mit der oberen Stufe Level 0, dessen Kategorien die Informationsbedürfnisse eher oberflächlich definieren. Die Kategorien in Level 1, welche auf die beiden im Level 0 aufgeteilt sind, sind hingegen präziser auf die vorkommenden Äußerungen abgestimmt. Entsprechend der Koch-Taxonomie fallen unter den Typ *Competence*, welchem 145 Äußerungen der Stichprobe zugeordnet sind, die Kategorien *Preparation* und *Manual Technique*, zu *Fact* hingegen zählen die sieben Label *Amount*, *Material*, *Instruction*, *Time*, *Equipment*, *Knowledge* und *Miscellaneous*, welche in diesem Teil des Datensatzes auf 82 Äußerungen aufgeteilt sind.

In Abbildung 3 sind die Verteilungen der Äußerungen pro Informationsbedürfnis im Level 1 aufgeführt. Die Kategorie *Preparation* wurde mit 138 (60.79%)

den meisten Äußerungen zugeteilt, *Knowledge* 27 (11.89%), *Instruction* 16 (7.05%), *Equipment* 13 (5.73%), *Material* 12 (5.29%), *Manual Technique* 7 (3.08%), *Miscellaneous* 5 (2.20%), *Amount* ebenfalls 5 (2.20%) und *Time* 4 (7.76%). Demnach ist *Preparation* mit 60.79% das mit Abstand am häufigsten vorkommende Informationsbedürfnis.

Merkmale und Unterschiede dieser beiden Level 0 Gruppen werden nun im Folgenden näher beschrieben.

4.1.4.1 Fact-basierte Informationsbedürfnisse

Äußerungen, welche dieser Kategorie zugeordnet sind, beziehen sich auf *Fact*-basierte Informationsbedürfnisse bezüglich Anweisungen während des Prozesses, die Fakten vermitteln. Einige dieser beinhalten häufig Schlüsselwörter, die auf die spezifische Kategorie hindeuten. In Äußerungen der Level 1 Kategorie *Amount* tauchen Wörter wie „often“, „much“ oder „regularly“ auf, welche auf ein mengenbezogenes Informationsbedürfnis hindeuten. Somit können Äußerungen wie „How often should I be watering my plant?“ (Konversation 11) oder „Could you specify how much cement I'll need?” (Konveration 14) einfach als Kategorie *Amount* ermittelt werden. Auch *Equipment* enthält Schlüsselwörter, beispielsweise „tools“, durch welche die Äußerung leicht der richtigen Kategorie zugeordnet werden kann wie in „What tool do I need to test the soil pH?“ (Konversation 17). Äußerungen, welche zum Informationsbedürfnis *Time* gehören, enthalten am Anfang häufig ein dafür typisches Schlüsselwort „long“, so auch in den Fragen „How long does it take for the basil to grow?“ (Konversation 2) und „Ok. Once thats done, how long should I let it dry for?” (Konversation 13).

Anhand solcher genannten Schlüsselwörter können *Fact*-basierte Informationsbedürfnisse leichter identifiziert werden, was eine große Erleichterung/Unterstützung für den Annotationsprozess darstellt.

4.1.4.2 Competence-basierte Informationsbedürfnisse

Äußerungen, welche dem Level 0 *Competence* zugeordnet sind, beziehen sich auf die Durchführung des Prozesses, also Handlungen oder Aktionen, welche für die Umsetzung des DIY-Projekts erforderlich sind. Ebenso wie einige Kategoiren

vom Level 0 *Fact*, zeigte *Preparation* häufige Schlüsselwörter auf. Dessen Äußerungen decken Informationsbedürfnisse zu den einzelnen Schritten des Prozesses ab, einschließlich Fragen zum nächsten Schritt oder Tipps, die während oder nach dem Projekt befolgt werden sollten. Ein enorm häufiges Schlüsselwort war demnach „next“. In Abbildung 3 ist zu sehen, dass diese Kategorie mit Abstand am häufigsten auftritt, wobei die meisten Äußerungen sich auf den nächsten Schritt des Projekts beziehen wie in „Gotcha. The tree is all ready for potting now, what's next?" (Konversation 1), "Okay, now what should I do next?" (Konversation 5) oder auch "What should my next step be?" (Konversation 7). *Manual Technique* hingegen enthält keine typischen Schlüsselwörter, weshalb die Einordnung in diese Kategorie herausfordernder ist. Äußerungen dessen sind Informationsbedürfnisse bezüglich einer praktischen Anleitung oder Methode wie ein Schritt genau auszuführen ist, welcher bisher noch nicht detailliert beschrieben wurde. Diesbezüglich war der Kontext und Gesprächsverlauf des laufenden Projekts notwendig, um den Prozess im Überblick zu behalten und einordnen zu können ob die Äußerung zu *Manual Technique* oder eher einer anderen Kategorie wie *Preparation* oder dem *Fact*-basierten Label *Knowledge* zugeteilt werden sollte. Falls die Situation oder auszuführende Aktion, auf die sich die Äußerung bezieht, bereits erklärt wurde, würde diese *Manual Technique* zugeordnet werden, andernfalls zu *Preparation*. Ein Beispiel dafür ist die Äußerung „I don't really have access to those right now but i have some beeswax lipbalm that is mostly beeswax. Where do I apply it to the zipper to unstick it?" (Konversation 3). Im vorherigen Verlauf des Gesprächs wurde bereits eine Anweisung gegeben, allerdings war diese nicht detailliert genug beschrieben, sodass der Teilnehmer dieses Gesprächs nach der genauen Ausführung fragte. Demnach wird diese Kategorie zu *Manual Technique* geordnet. Dieses Beispiel illustriert eine Herausforderung während des Annotationsprozesses.

Durch die beschriebenen Herausforderungen wird deutlich, dass ein Sprachmodell, welches im Bereich Conversational Search eingesetzt wird, ein sehr gutes Kontextverständnis benötigt. Hierfür bieten sich die LLMs GPT-3 und GPT-4 an,

welche durch ihre Transformer-Architektur ein umfassendes Verständnis der natürlichen Sprache und zudem eine starke Leistungsfähigkeit aufweisen.

4.2 Datenannotation mit GPT-Modellen

Da die manuelle Annotation mit der neuen Taxonomie abgeschlossen ist, wird nun untersucht ob diese erstellte Taxonomie und dadurch resultierende Stichprobe der Daten ausreichen, damit ein LLM die Informationsbedürfnisse korrekt identifizieren kann. Hierfür wird mit verschiedenen Prompting-Strategien GPT-3 und GPT-4 für die Datenannotation genutzt. Demnach stellt sich folgende Frage: Wie gut können die GPT-Modelle die Informationsbedürfnisse im Bereich DIY vorhersagen?

4.2.1 Datensatz

Kategorie	**Stichprobe**	**Train-Datensatz**	**Test-Datensatz**
Equipment	13	4	9
Preparation	138	68	70
Material	12	5	7
Miscellaneous	5	1	4
Instruction	16	6	10
Knowledge	27	15	12
Manual Technique	7	2	5
Time	4	1	3
Amount	5	2	3

Tabelle 2: Verteilungen der Level 1 Informationsbedürfnisse in der Stichprobe, sowie im Train- und Test-Datensatz

Die Stichprobe wurde den Modellen als Datensatz übergeben, wobei diese auf Grund der für das *Few-Shot* benötigten Trainingsdaten in einen Train- und einen Test-Datensatz aufgeteilt wurden. Insgesamt umfasst der Datensatz $N = 456$ Äußerungen, wobei nur 227 *Student* angehören, die restlichen 229 stammen von *Teacher* und wurden nur für das Kontextverständnis übergeben, damit die Konversationen als Ganzes betrachtet werden konnten und keine zusammenhangslosen

Äußerungen aufeinanderfolgten. In Tabelle 2 wird die Aufteilung der beiden Datensätze gezeigt. Bei der Aufteilung wurde darauf geachtet, dass die sehr selten vorkommenden Informationsbedürfnisse in beiden Datensätzen auftreten. Dadurch, dass der Kontext mit einbezogen wurde, konnten die Konversationen nicht aufgelöst werden, damit diese als Ganzes betrachtet werden können. Somit konnten die Kategorien nicht gleichmäßig aufgeteilt werden, da sie nicht gleich häufig in den Konversationen auftreten. Dadurch beinhaltete der Test-Datensatz insgesamt 123 Äußerungen von *Student*, welche letztlich von den beiden Sprachmodellen annotiert wurden.

Neben den Äußerungen beinhaltet der Datensatz noch die Zuordnungen der Teilnehmer zu *Student* oder *Teacher*, da diese Informationen bei der Datenannotation benötigt werden, um unterscheiden zu können welche zugehörigen Äußerungen klassifiziert werden und welche nur für den Kontext dienen.

4.2.2 Prompting

Sowohl das *Zero-Shot*, als auch das *Few-Shot* Prompting wird angewandt, da sich diese beiden Strategien gut dafür eignen, wenn keine große Datenmenge zur Verfügung steht. Für die Annotation der Äußerungen bezüglich Level 0 wird dem LLM ein Prompt übergeben, mit Hilfe dessen die Daten mit einer der beiden Kategorien *Fact* oder *Competence* klassifiziert wird. Bei der Klassifikation der Level 1 Kategorien wird zwischen den neun verschiedenen Informationsbedürfnissen unterschieden. Diese beiden Strategien werden im Nachhinein verglichen und diskutiert, um die bessere Methode

Damit die beiden autoregressiven Sprachmodelle die Klassifizierungsmerkmale pro Kategorie verstehen, wurde iterativ ein Prompt erzeugt, was in Abschnitt 3.3.1.2 detailliert beschrieben wurde. Im Folgenden werden die Ergebnisse der beiden GPT Modelle beschrieben und diskutiert.

4.2.3 GPT-3 im Vergleich zu GPT-4

Modell	Strategie	Weighted Avg F1	Cohen's Kappa
GPT-3	Zero-Shot	0.82	0.63
	Few-Shot	0.91	0.81

GPT-4	Zero-Shot	0.97	0.93
	Few-Shot	0.98	0.95

Tabelle 3: Vergleich der Ergebnisse bei Level 0, gruppiert nach Modell und Prompting-Strategie

Modell	Strategie	Weighted Avg F1	Cohen's Kappa
GPT-3	Zero-Shot	0.76	0.64
	Few-Shot	0.86	0.79
GPT-4	Zero-Shot	0.91	0.86
	Few-Shot	0.93	0.90

Tabelle 4: Vergleich der Ergebnisse bei Level 1, gruppiert nach Modell und Prompting-Strategie

Insgesamt erzielte GPT-4 durchgehend bessere Ergebnisse als GPT-3. Um dies zu messen, wurden die bei der Textklassifikation typischen Metriken verwendet, sodass sich bei jedem Modell pro Prompting-Strategie ein *Weighted Average F1* (gewichteter Durchschnitt des *F1-Scores*) ergab. Durch *Cohen's Kappa* wurde die Übereinstimmung zwischen dem Modell mit der Ground Truth, also den manuellen Labeln gemessen, unter Berücksichtigung eines zufälligen Ergebnisses bei der Kategorisierung.

Sowohl bei Level 0, als auch bei Level 1 stieg beim *Zero-Shot* Prompting sowohl der Wert des *Weighted Average F1* (gewichteter Durchschnitt des *F1-Scores*) um 0.15, als auch der *Cohen's Kappa* Wert durchschnittlich um 0.26. Nicht ganz so groß waren die Unterschiede zwischen den beiden Modellen beim *Few-Shot* Prompting, aber dennoch aussagekräftig. Der *Weighted Average F1*-Wert stieg, sowohl bei Level 0, als auch Level 1, der *Cohen's Kappa* Wert um 0.125 im Durchschnitt.

Auf Grund dessen werden im Folgenden nur die Ergebnisse von GPT-4 beschrieben, da dieses Modell genauere Ergebnisse bei der Klassifikation erzielte und deswegen dieses Modell letztlich für die darauffolgende Annotation des gesamten Datensatzes genutzt wird.

4.2.4 Ergebnisse der verschiedenen Prompting-Strategien mit GPT-4 und deren Diskussion

Durch *Precision (P), Recall (R)* und dem *F1-Score (F1)* werden für jede einzelne Kategorie die Leistung des Modells bezüglich der Genauigkeit (P), Erkennungsrate *(R)* und deren harmonischem Mittel *(F1)* gezeigt. Die *Weighted Average* dieser Metriken bieten einen Gesamteindruck über alle Kategorien zusammen, wobei die Verteilung der Kategorien berücksichtigt wird. Durch den *Cohen's Kappa* Wert wird die Übereinstimmung zwischen der Ground Truth und den vom Modell generierten Labels gemessen, wobei die Wahrscheinlichkeit einer zufälligen Übereinstimmung berücksichtigt wird.

4.2.4.1 Ergebnisse Level 0

In den Tabellen Tabelle 5 und Tabelle 6 sind die Werte der Metriken *Precision, Recall, F1-Score* sowie die zugehörigen *Weighted Averages* aufgeführt. Die Werte zwischen dem *Zero-* und *Few-Shot* Prompting unterscheiden sich kaum, nur die *Precision* beim Level 0 Typ *Competence* steigt um 0.01 und der *Recall* von *Fact* um 0.02, sowie die *Weighted Averages* um 0.01. Somit zeigt das Prompting mit der Verwendung von Beispieldaten ein besseres Ergebnis auf. Der *Cohen's Kappa* Wert beträgt beim *Zero-Shot* 0.93, bei *Few-Shot* stieg er auf 0.95. Dennoch gelten beide Werte der verschiedenen Methoden als fast perfektes Ergebnis (Landis & Koch, 1977).

	Precision	**Recall**	**F1-Score**
Fact	0.98	0.94	0.96
Competence	0.96	0.99	0.97
Weighted Avg	0.97	0.97	0.97

Tabelle 5: Ergebnisse GPT-4 Level 0 bei Zero-Shot Prompting

	Precision	Recall	F1-Score
Fact	0.98	0.96 +0.02	0.97 +0.01
Competence	0.97 +0.01	0.99	0.98 +0.01
Weighted Avg	0.98 +0.01	0.98 +0.01	0.98 +0.01

Tabelle 6: Ergebnisse GPT-4 Level 0 bei Few-Shot Prompting (mit der Differenz zu Zero-Shot Promting)

4.2.4.2 Ergebnisse Level 1

In den Tabellen Tabelle 7 und Tabelle 8 sind die Ergebnisse der Metriken aufgeführt. Die Informationsbedürfnisse *Equipment* und *Amount* zeigen bei beiden Prompting-Strategien perfekte Ergebnisse, obwohl letztere sogar nur ein Trainingsbeispiel besitzt. Die trotzdem perfekten Werte lassen sich mit den häufig auftretenden Schlüsselwörtern begründen (siehe Kapitel 4.1.4.1). Auch *Miscellaneous* und *Manual Technique* weisen gleichbleibende Werte auf: zwar sind die *Precision*-Werte mit 1.00 perfekt, die Werte für *Recall* und *F1-Score* sind jedoch die schlechtesten aller Kategorien. Wie bereits in Kapitel 4.1.4.2 erwähnt, bringt die Kategorisierung von *Manual Technique* einige Herausforderungen mit sich, da es keine klaren Schlüsselbegriffe gibt und der Kontext beziehungsweise die Historie der Konversation eine große Rolle spielt. Auch *Miscellaneous* besitzt keine typischen und häufig vorkommenden Begriffe. Diesem Problem, dass diese nicht klar identifizierbar sind, könnten Trainingsbeispiele entgegenwirken. Da diese beiden Kategorien allerdings relativ selten vorkommen, war auch deren Anteil an dem Train-Datensatz extrem gering. Dies trägt nicht nur dazu bei, dass die Informationsbedürfnisse nicht klar erkannt wurden, sondern auch, dass sich die Werte zwischen *Zero-* und *Few-Shot* nicht änderten, da so eine geringe Anzahl an Beispielen für solche Kategorien nicht ausreichend ist. Obwohl auch bei *Time* nur ein einziges Trainingsbeispiel vorhanden ist, erreichte das Informationsbedürfnis durch das *Few-Shot* Prompting einen *Recall* von 1.00 und weist dadurch in allen drei Metriken perfekte Ergebnisse auf. Dies deutet darauf hin, dass für diese Kategorie nicht unbedingt viele Beispieldaten notwendig sind, da sie trotzdem aussagekräftige Schlüsselwörter besitzt (siehe Kapitel 4.1.4.1). *Preparation* verbesserte sich mit 0.03 mehr in *Precision* und mit 0.01 mehr bei *F1-Score* leicht bei *Few-Shot*,

wobei die Ergebnisse beim *Zero-Shot* schon nahezu perfekt waren. Diese Kategorie enthält sehr häufige Schlüsselwörter, weshalb diese durchgehend deutlich identifizierbar ist. *Material* verbesserte sich durch das *Few-Shot* Prompting in allen drei Metriken um durchschnittlich 0.16. Auch *Knowledge* verbesserte sich leicht in Precision um 0.04. *Instruction* ist die einzige Kategorie, die sich durch die Strategie mit Beispieldaten in einem Wert verschlechterte. Dadurch, dass einer zusätzlichen Äußerung statt dem richtigen das falsche Label *Instruction* zugeordnet wurde, was demnach ein *False Positive* ist, sank *Precision* um 0.03, *Recall* verbesserte sich jedoch um 0.10, sodass auch der *F1-Score* leicht stieg.

Alle *Weigthed Averages* stiegen leicht an, *Precision* um 0.03, *Recall* und *F1-Score* um 0.02. Demnach ist das *Few-Shot* Prompting insgesamt besser als das *Zero-Shot* Prompting. Außerdem beträgt der *Cohen's Kappa* Wert beim *Zero-Shot* Prompting 0.86, bei *Few-Shot* 0.90. Wie auch bei den Vorhersagen der Level 0 Kategorien zeigen also beide Prompting-Strategien ein fast perfektes Ergebnis auf (Landis & Koch, 1977).

Um zu testen ob die beobachteten Ergebnisse beziehungsweise Unterschiede statistisch signifikant sind, wurden folgende Hypothesen für die jeweiligen Metriken *Precision*, *Recall* und *F1-Score* untersucht:

- Nullhypothese (*H0*): Es gibt keinen signifikanten Unterschied in der Metrik zwischen der *Zero-* und der *Few-Shot* Strategie
- Alternativhypothese (*H1*): Es gibt einen signifikanten Unterschied in der Metrik zwischen der *Zero-* und der *Few-Shot* Strategie

Um die Hypothesen zu überprüfen, wurde ein Wilcoxon Signed-Rank Test durchgeführt in Verbindung mit einer Bonferroni-Korrektur, um das Risiko von falsch-positiven Ergebnissen zu minimieren. Da Level 0 nur zwei Kategorien besitzt und demnach zwei Beobachtungen pro Metrik, wurde hierfür kein statistischer Test durchgeführt, da die Anzahl zu gering ist.

In Bezug auf *Precision* zeigte der Test keinen statistisch signifikanten Unterschied zwischen der *Zero-* und *Few-Shot* Strategie ($p > 0.0167$), was darauf hindeutet, dass beide eine vergleichbar gute Genauigkeit besitzen. Auch für *Recall* gab es zwischen den beiden Strategien keinen signifikanten Unterschied ($p > 0.0167$).

Somit identifizieren beide Methoden relevante Ergebnisse in einem vergleichbaren Umfang. Auch bei der dritten Metrik *F1-Score* zeigte sich zwischen *Zero-* und *Few-Shot* Prompting kein signifikanter Unterschied ($p > 0.0167$). Das bedeutet, dass es keinen statistischen Beweis dafür gibt, dass das *Few-Shot* Prompting in den drei Metriken deutlich besser ist als *Zero-Shot*. Diese Ergebnisse könnten darauf zurückzuführen sein, dass für jede Metrik nur neun Beobachtungen vorhanden waren, was als kleine Stichprobe angesehen wird und es dadurch wahrscheinlicher ist, dass Unterschiede, welche exisitieren, trotzdem nicht erkannt werden.

	Precision	**Recall**	**F1-Score**
Equipment	1.00	1.00	1.00
Preparation	0.97	0.99	0.98
Material	0.83	0.71	0.77
Miscellaneous	1.00	0.50	0.67
Instruction	0.70	0.70	0.70
Knowledge	0.71	1.00	0.83
Manual Technique	1.00	0.60	0.75
Time	1.00	0.67	0.80
Amount	1.00	1.00	1.00
Weighted Avg	0.92	0.91	0.91

Tabelle 7: Ergebnisse GPT-4 Level 1 bei Zero-Shot Prompting

	Precision	**Recall**	**F1-Score**
Equipment	1.00	1.00	1.00
Preparation	1.00 +0.03	0.99	0.99 +0.01
Material	1.00 +0.17	0.86 +0.15	0.92 + 0.15
Miscellaneous	1.00	0.50	0.67
Instruction	0.67 -0.03	0.80 +0.10	0.73 +0.03
Knowledge	0.75 +0.04	1.00	0.86 +0.03
Manual Technique	1.00	0.60	0.75
Time	1.00	1.00 +0.33	1.00 +0.20

Amount	1.00	1.00	1.00
Weighted Avg	0.95 +0.03	0.93 +0.02	0.93 +0.02

Tabelle 8: Ergebnisse GPT-4 Level 1 bei Few-Shot Prompting (mit der Differenz zu Zero-Shot Promting)

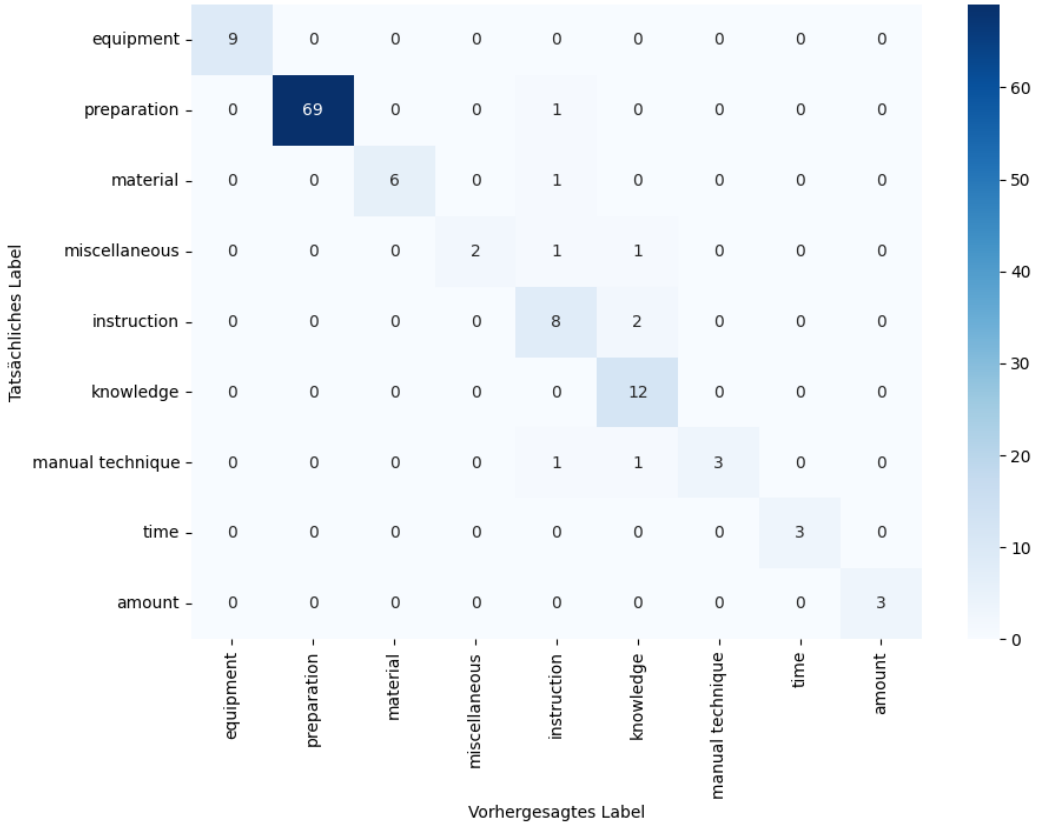

Abbildung 4: Konfusionsmatrix Few-Shot Prompting Level 1

Obwohl die aufgezeigten Ergebnisse statistisch gesehen nicht signifikant sind, ist das *Few-Shot* Prompting überzeugender, sowohl auf Grund der Verbesserungen der Metriken, als auch wegen der Erhöhung des *Cohen's Kappa* Werts. Dies zeugt also von einer höheren Überstimmung mit der manuellen Annotation. Dass die Vorhersagen des Modells mit dieser Strategie fast bei jedem Informationsbedürfnis übereinstimmten, ist in Abbildung 4 nochmals visuell dargestellt.

Somit wird diese Prompting-Strategie für die Annotation von Level 1 für den kompletten Datensatz der Wizard-of-Tasks Studie angewandt. GPT-4 wies gute Ergebnisse auf, wodurch davon ausgegangen werden kann, dass die Informationsbedürfnisse verlässlich annotiert werden. Somit kann eine große Menge an domänenspezifischen Daten gewonnen werden, was eine Voraussetzung für die

Entwicklung von Conversational Assistants für solche prozeduralen Bereiche wie DIY ist, damit ein Sprachmodell für diese Anwendung trainiert werden kann.

4.3 Ergebnisse der Annotation durch GPT-Modelle

Mit Hilfe von GPT-4 und dem *Few-Shot* Prompting wurde der komplette Wizard-of-Tasks Datensatz gelabelt beziehungsweise die Äußerungen von *Student*, welcher den Nutzer des Conversational Assistants imitierte. Hierbei wurde aber nur Level 1 betrachtet, da diese die detaillierteren Infromationsbedürfnisse darstellen. Zudem kann mit diesem erstellten Datensatz die Verteilung zwischen den beiden Level 0 Kategorien herausgefiltert werden.

Zunächst folgt eine kurze Beschreibung des Datensatzes, welcher für die Annotation übergeben wurde. Anschließend werden die Ergebnisse des neuen Datensatzes aufgezeigt.

4.3.1 Datensatz

Zunächst wurden aus dem Datensatz alle Äußerungen entfernt, welchen der Intent *stop* zugeordnet war. Da die Nutzer mit ihrer Äußerung die Konversation stoppen wollen und keine Antwort mehr vom Conversational Assistant erwarten, gelten diese nicht als Informationsbedürfnis. Zusätzlich wurden alle Zeilen des Datensatzes entfernt, bei denen *turn* leer ist, da ansonsten ein Fehler bei der Annotation entstehen könnte. Somit besteht der Datensatz, der GPT-4 für die Annotationsaufgabe übergeben wurde, aus 4928 Äußerungen seitens *Teacher*, welche für das Kontextverständnis dienen und 4887 (knapp 50%) seitens *Student*, welche letztlich annotiert werden.

4.3.2 Annotationsergebnisse

Bis auf sieben Äußerungen, welche mit zwei Informationsbedürfnissen klassifiziert wurden, wurden alle Äußerungen richtig mit nur einer Kategorie versehen. Da die Studie jedoch Multiklassen-Textklassifikation behandelt, werden die Äußerungen mit zwei zugeteilten Kategorien manuell überprüft und anhand des Kontexts die passendere beziehungsweise relevantere Kategorie gewählt, um den Datensatz trotzdem gut analysieren zu können.

Schließlich entsteht dadurch ein Datensatz dessen Kategorien durchschnittlich 543 Äußerungen zugeteilt sind (*min* = 99.00, $x_{.25}$ = 193.00, $x_{.50}$ = 237.00, $x_{.75}$ = 369.00, *max* = 2593.00, *sd* = 793.44). Durchschnittlich hatte eine Äußerung von *Student* eine Länge von 13.58 Wörtern (*min* =3.0, *max* = 56.0).

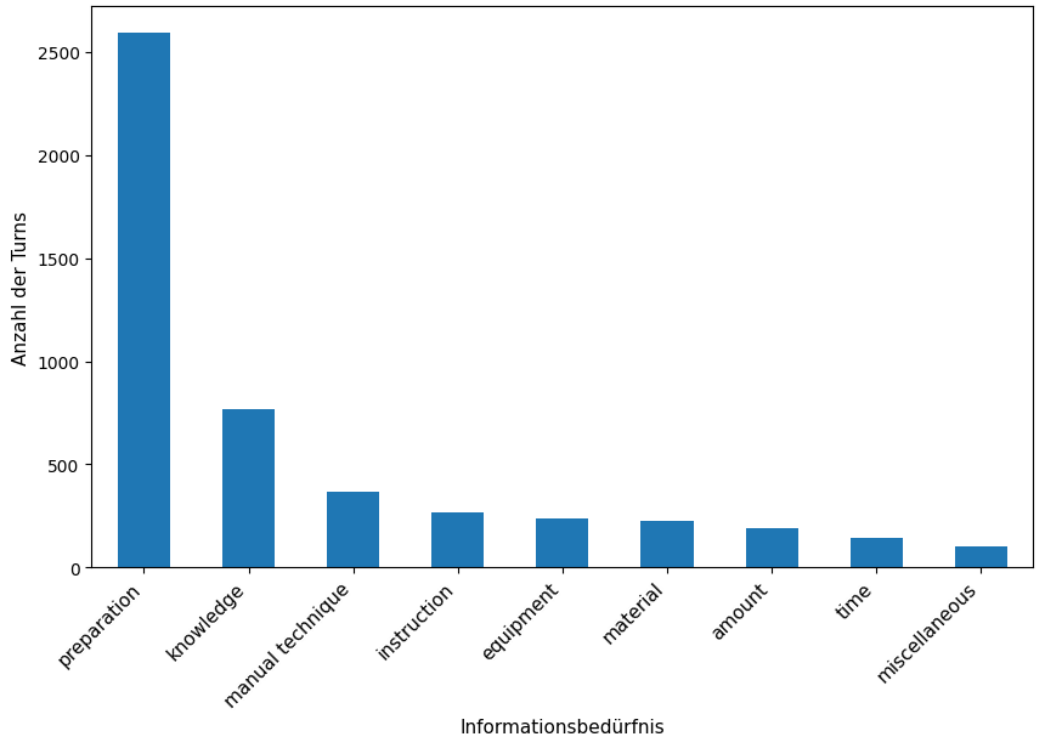

Abbildung 5: Verteilungen der der Informationsbedürfnisse des gesamten Datensatzes Wizard-of-Tasks

In Abbildung 5 sind die Verteilungen der Äußerungen pro Informationsbedürfnis im Level 1 aufgeführt. Die Kategorie *Preparation* wurde mit 2593 (53.06%) den meisten Äußerungen zugeteilt, *Knowledge* 766 (15.67%), *Manual Technique* 369 (7.55%), *Instruction* 265 (5.42%), *Equipment* 237 (4.85%), *Material* 224 (4.58%), *Amount* 193 (3.95%) , *Time* 141 (2.89%) und *Miscellaneous* 99 (2.03%). Demnach ist *Preparation* mit knapp über der Hälfte der Äußerungen das mit Abstand am häufigsten vorkommende Informationsbedürfnis, was auch bei der Stichprobe der Fall ist (siehe Abschnitt 4.1.4.1). Insgesamt sind dem Level 0 Typ *Competence* demnach 2962 (60.60%) und *Fact* 1925 (39.40%) Äußerungen zugeteilt.

5 Diskussion

Im Folgenden werden die Ergebnisse der Annotation der Stichprobe mit denen des gesamten Datensatzes verglichen und diskutiert, um dadurch herauszufinden ob die relativ kleine Stichprobe repräsentativ genug dafür ist, um durch *Few-Shot* Prompting mit GPT-4 eine große relevante Datenmenge zu erzeugen, was die zweite Forschungsfrage addressiert. Anschließend wird die Taxonomie von Frummet et al. (2022) mit der neu erstellten im Bereich DIY verglichen, um damit Schlussfolgerungen bezüglich der Übertragbarkeit auf eine andere Domäne und somit Generalisierbarkeit dieser Taxonomie zu ziehen, was die zentrale Forschungsfrage dieser Arbeit darstellt.

5.1 Ergebnisse der Annotation des gesamten Datensatzes

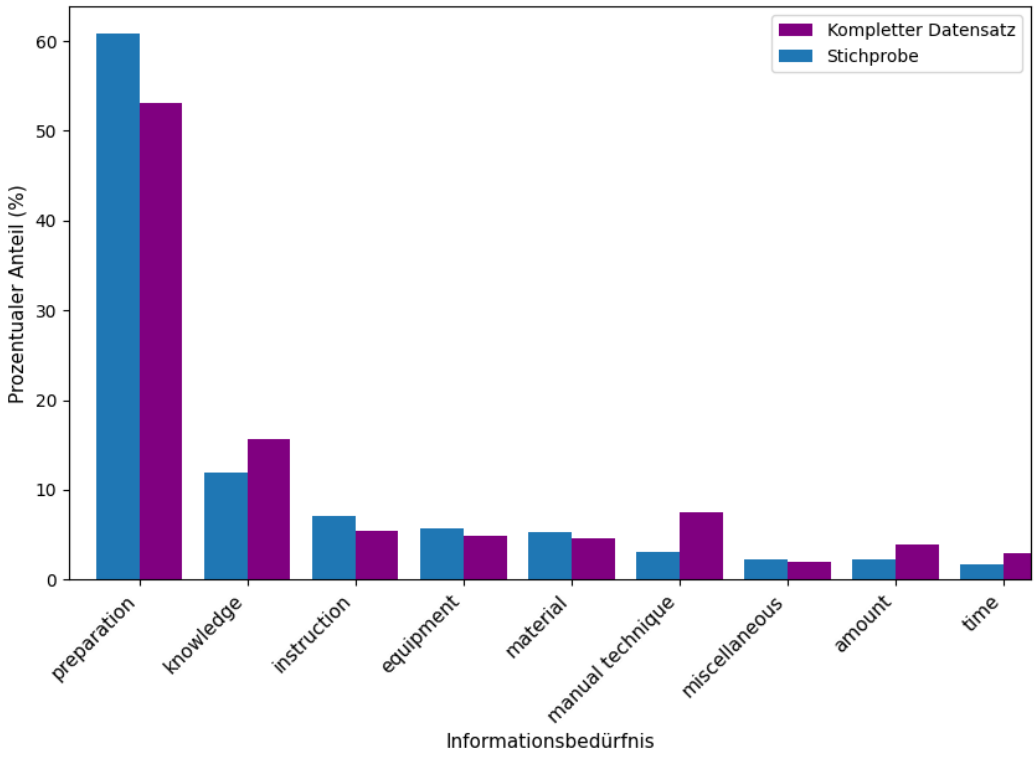

Abbildung 6: Prozentualer Anteil der Informationsbedürfnisse gruppiert nach Datensatz

Die Verteilungen der Informationsbedürfnisse verglichen zwischen der Stichprobe und dem gesamten Datensatz scheint relativ ähnlich zu sein, was Abbildung 6 zu sehen ist. Um dies statistisch zu prüfen, werden die folgenden beiden Hypothesen untersucht:

- Nullhypothese (*H0*): Es gibt keinen signifikanten Unterschied in den Verteilungen der Kategorien zwischen der Stichprobe und dem vollständigen Datensatz
- Alternativhypothese (*H1*): Es gibt einen signifikanten Unterschied in den Verteilungen der Kategorien zwischen der Stichprobe und dem vollständigen Datensatz

Mit Hilfe des Chi-Quadrat-Tests wurde ermittelt, dass es keinen statistisch signifikanten Unterschied zwischen den beiden Verteilungen gibt ($p > 0.05$), wodurch *H0* nicht abgelehnt werden kann. Somit kann angenommen werden, dass die Stichprobe repräsentativ für den kompletten Datensatz ist, da die Verteilungen der Kategorien ähnlich genug sind, um dadurch anzunehmen, dass die Erkenntnisse aus der Stichprobe übertragbar sind.

Demnach stellt die Stichprobe trotz ihrer geringen Größe eine gute Grundlage dar. Zusammen mit dem *Few-Shot* Prompting mit GPT-4 kann eine große Menge an relevanten Annotationsdaten geschaffen werden, was den Annotationsprozess erheblich erleichtert. Somit ist es möglich mit Hilfe einer Annotationsvorlage und einem LLM eine große Menge an annotierten Daten zu schaffen, womit die dritte Forschungsfrage beantwortet wird.

5.2 Vergleichende Analyse der beiden Taxonomien

Zunächst werden die Gemeinsamkeiten und Unterschiede der beiden Taxonomien analysiert. Dabei werden die Ähnlichkeiten der beiden Domänen einbezogen und erläutert. Durch diese Analyse wird eine mögliche Übertragbarkeit auf andere prozedurale Bereiche und somit Generalisierbarkeit der Taxonomie von Frummet et al. (2022) diskutiert.

5.2.1 Übertragung der Koch-Taxonomie auf DIY

Level 1 Kochen	**Level 1 DIY**	**Übertragbarkeit**
Amount	Amount	Ja
Ingredient	Material	Ja, mit Umformulierung
Preparation	Preparation	Ja
Cooking Technique	Manual Technique	Ja, mit Umformulierung

Recipe	Instruction	Ja, mit Umformulierung
Time	Time	Ja
Equipment	Equipment	Ja
Knowledge	Knowledge	Ja
Meal	-	Nein
Temperature	-	Nein
Miscellaneous	Miscellaneous	Ja

Tabelle 9: Label der beiden Level 1 in der Domäne Kochen und DIY

Insgesamt konnte die Koch-Taxonomie mit einigen Anpassungen auf den Bereich DIY gut übertragen werden. Die beiden Kategorien *Fact* und *Competence,* welche der obersten Ordnung Level 0 zugeordnet sind, wurden direkt übernommen. Der Prozess des Kochens ist genauso aufgebaut wie der Prozess eines DIY-Projekts, da sich Nutzer nach den benötigten Arbeitsmitteln und Fakten dazu, sowie dem Ablauf und Durchführung des Prozesses erkundigen, weshalb diese beiden Label zu beiden Domänen passen.

Die Labels der zweiten Ordnung sind jedoch etwas spezifischer. Bei diesen wurden zunächst die domänenspezifischen Begriffe des Kochens an den DIY-Kontext angepasst. Somit wurde aus der Kategorie *Ingredient* die Kategorie *Material*, aus *Cooking Technique* entstand *Manual Technique*, *Recipe* wurde zu *Instruction* umformuliert und *Meal* zu *Task*. Die restlichen Kategorien *Amount*, *Preparation, Time, Equipment, Knowledge, Temperature und Miscellaneous* sind domänenoffene Bezeichnungen und somit neben der Domäne Kochen ohne Umformulierungen auch für einen anderen Bereich passend.

Durch den iterativen Prozess der manuellen Annotation der Stichprobe wurden sowohl die Taxonomie, als auch die Definitionen der Label nach und nach angepasst. Dadurch blieben von den anfangs elf aufgestellten Kategorien nur noch neun übrig. Die Kategorie *Temperature* wurde in der Stichprobe für keine Äußerung als passende Kategorie empfunden, weshalb diese aussortiert wurde. Demnach ist diese Kategorie eher domänenspezifisch und nicht auf andere Themenbereiche wie DIY übertragbar. Neben *Temperature* wurde ebenso die Kategorie *Task* aussortiert, da diese nicht klar von den anderen Kategoiren abgegrenzt

werden konnte. Vor allem zur Kategorie *Instruction* konnte keine klare Abgrenzung definiert werden, weshalb der Prozess der Annotation mit dieser deutlich erschwert werden würde.

Die entstandene Taxonomie für die Domäne DIY besteht aus den beiden Level 0 Typen *Fact* und *Competence*, sowie den neun Level 1 Informationsbedürfniskategorien *Amount*, *Material*, *Preparation*, *Manual Technique*, *Instruction*, *Time*, *Equipment*, *Knowledge* und *Miscellaneous*. So wie im Bereich Kochen nach möglichen Rezepten gesucht wird, werden hier mögliche Projekte oder allgemeine Informationen zu DIY-Projekten erfragt, weshalb die Kategorie *Instruction* (vorher *Recipe*) in dieser Domäne passend ist. Zudem werden Informationen zu Arbeitsmitteln und zugehörige Fakten wie Mengen oder Zeitangaben benötigt, weshalb die Informationsbedürfnisse *Material* (vorher *Ingredient*) nach kurzer Kontextanpassung sowie *Amount*, *Equipment* und *Time* in der Taxonomie aufgenommen werden konnten. Fragen nach den einzelnen Schritten tauchen in jedem prozeduralen Bereich auf, weshalb *Preparation* ebenso übernommen werden konnte, genauso wie Fragen bezüglich Definitionen oder Hintergrundwissen, wodurch *Knowledge* auch in die neue Taxonomie übernommen wurde. Da der Bereich DIY genauso wie der Bereich Kochen sehr spezifisch ist, können Situationen auftreten, in denen bestimmte Techniken oder Ausführungen benötigt werden, die ein gewisses Vorwissen benötigen. Deshalb wurde nach Umformulierung auf den Kontext passend *Manual Technique* (vorher *Cooking Technique*) auf die neue Taxonomie übertragen. Zuletzt wurde die Kategorie *Miscellaneous* übernommen für Äußerungen, welche den anderen Kategorien nicht zugeordnet werden konnten. Um einen besseren Überblick zu gewähren, werden in der Tabelle 9 die Kategorien der beiden Domänen gegenübergestellt mit einem kurzen Statement der Übertragbarkeit.

Durch Tabelle 9 ist zu erkennen, dass sich die Taxonomie aus dem Bereich Kochen gut auf den Bereich DIY übertragen lässt. Dadurch, dass diese beiden prozeduralen Bereiche gleich aufgebaut sind, können mit einigen Anpassungen an den Kontext fast alle Kategorien übertragen werden, was demnach für eine gute Übertragbarkeit der Taxonomie-Vorlage spricht.

5.2.2 Verteilungen der Informationsbedürfnisse der Taxonomien

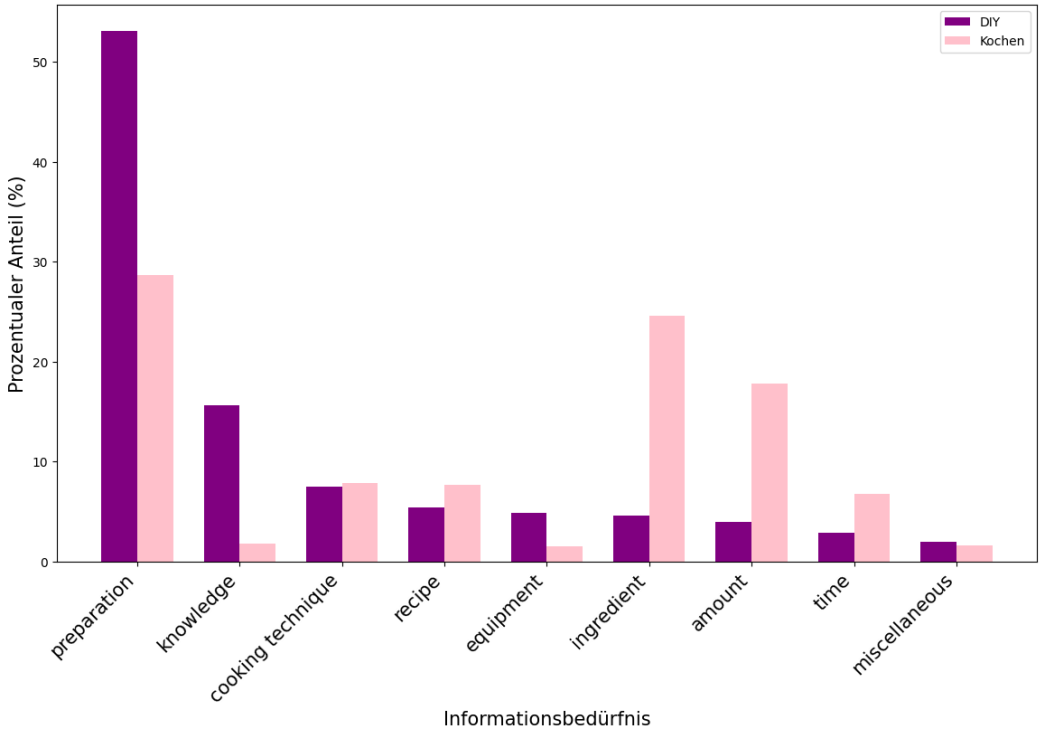

Abbildung 7: Prozentualer Anteil der Informationsbedürfnisse gruppiert nach Datensatz (die angepassten DIY-Kategorien wurden hierfür der zugehörigen Koch-Kategorie zugeordnet)

Um die Verteilungen der Kategorien der beiden Taxonomien analysieren zu können, wurden hierfür DIY-Informationsbedürfnisse denen der Koch-Taxonomie zugeordnet, von denen sie abstammen (siehe 5.2.1). In der Abbildung 7 sind die Verteilungen der Kategorien der beiden Taxonomien gezeigt.

Preparation ist mit über 50% die am häufigsten vertretene Kategorie in der DIY-Taxonomie, wohingegen der Anteil in der anderen Domäne deutlich geringer ist. Ein möglicher Grund dafür könnten umfangreichere Vorbereitungsschritte bei einem DIY-Projekt sein, wodurch diese Kategorie einen zentralen Bereich ausmacht. Auch *Knowledge* tritt bei DIY viel häufiger auf, was darauf hinweist, dass Hintergrundwissen in diesem Bereich häufiger benötigt wird und demnach dieser Bereich komplexer als Kochen sein könnte. *Cooking Technique*, was in der Domäne DIY *Manual Technique* entspricht, ist in beiden Taxonomien ungefähr gleich verteilt. Auch *Recipe* (in DIY *Instruction*) weist einen relativ ähnlichen Anteil auf, wobei diese Kategorie im Bereich Kochen häufiger vorkommt.

Die könnte darauf hindeuten, dass sich in diesem Berich mehr Inspiration bezüglich möglicher Gerichte eingeholt wird, wohingegen im Bereich DIY das Projekt oftmals im Vorhinein schon feststeht. Die Kategorie *Equipment* ist im Bereich Kochen viel weniger vertreten als im DIY, was darauf hindeutet, dass in letzterer eher komplexere Projekte durchgeführt werden, mit deren Situationen die Nutzer nicht vertraut sind. Sowohl *Ingredient* (in DIY *Material*), als auch *Amount* sind in der Koch-Taxonomie zwei dominierende Kategorien, was darauf hinweist, dass die Wahl und auch Menge der Zutaten sehr im Fokus stehen, bei DIY hingegen scheinen Materialien und deren Menge nicht zu sehr im Fokus zu stehen. Die Kategorie *Time* ist in der Koch-Taxonomie ebenfalls öfter vertreten als bei DIY. Während des Kochens werden oftmals Zetangaben benötigt, bei DIY hingegen scheint Zeit kein wichtiger Faktor zu sein, um ein Projekt durchzuführen. *Miscellaneous* ist in beiden Kategorien relativ gleich vertreten, gesamt macht sie jedoch nur einen kleinen Teil aus. Das deutet darauf hin, dass nur wenige Äußerungen während der Prozesse geäußert werden, welche eher weniger relevant für den Prozess sind.

Preparation ist das meist dominierende Informationsbedürfnis im Bereich DIY, gefolgt von der Kategorie *Knowledge*. Das deutet darauf hin, dass der Fokus in diesem Bereich auf der Durchführung und diesbezüglichem Hintergrundwissen des Projekts liegt. Die dominierenden Informationsbedürfnisse im Bereich Kochen sind *Preparation*, *Ingredient* und *Amount*. Der Fokus liegt also auf den Zutaten und Mengenangaben dieser, sowie der Zubereitung des Gerichts.

Um die erläuterten Gemeinsamkeiten und Unterschiede quantitativ zu belegen, wurde der Pearson-Korrelationskoeffizient zwischen den Verteilungen der Kategorien der beiden Taxonomien ermittelt. Der Korrelationswert r=0.58 zeigt, dass es sowohl Ähnlichkeiten, als auch Unterschiede in den Verteilungen der Kategorien gibt. Dies bedeutet, dass manche Kategorien in beiden Taxonomien eine ähnliche Verteilung aufweisen, andere jedoch nicht, was bereits in 5.2.1 beobachtet wurde. Dies könnte darauf zurückzuführen sein, dass sich die beiden Bereiche auf Grund ihrer prozeduralen Struktur zwar ähnlich sind, jedoch nicht gleich.

Um dies statistisch zu prüfen, werden die folgenden beiden Hypothesen untersucht:

- Nullhypothese (*H0*): Es gibt keinen signifikanten Unterschied in der Verteilung der Kategorien zwischen der DIY-Taxonomie und der Koch-Taxonomie
- Alternativhypothese (*H1*): Es gibt einen signifikanten Unterschied in der Verteilung der Kategorien zwischen den beiden Taxonomien.

Durch Anwendung des Qui-Quadrat-Tests wurde ermittelt, dass *H0* verworfen und *H1* angenommen wird ($p < 0.05$). Die Unterschiede in den Verteilungen sind also statistisch signifikant, was darauf zurückzuführen sein könnte, dass in den beiden Domänen der Fokus bei verschiedenen Informationsbedürfnissen liegt.

Diese Unterschiede deuten darauf hin, dass die beiden Domänen verschiedene Schwerpunkte bezüglich der Informationsbedürfnisse aufweisen. Die Taxonomie von Frummet et al. (2022) ist demnach flexibel und anpassbar an ähnliche Bereiche ist, welche eine gleiche Struktur bezüglich des Ablaufs des Prozesses aufweisen. Dies verdeutlicht die Generalisierbarkeit der Taxonomie für prozedurale Bereiche.

6 Limitationen

Die Methode mit welcher der Datensatz Wizard-of-Tasks gesammelt wurde, behandelt zwar die Domäne DIY, imitiert jedoch keine reale Situation zwischen einem Nutzer und einem Conversational Assistant. Dadurch, dass die behandelten DIY-Projekte nicht ausgeführt, sondern nur besprochen wurden, gehen möglicherweise Äußerungen und damit Probleme unter, welche während der Durchführung auftreten könnten. Eine weitere Limitation dieser Arbeit ist die Anzahl der manuell annotierten Daten. Da auf Grund des Few-Shot Promptings die Daten aufgeteilt werden mussten, beinhaltete der Testdatensatz, der von den Large Language Models gelabelt wurde, nur 123 Äußerungen von *Student*. Einigen Kategorien waren dadurch weniger als zehn Äußerungen zugeteilt, sodass nicht erkenn-

bar ist ob die GPT-Modelle diese Kategorien gut generalisieren können. Eine weitere Limitation ist die Art der Textklassifikation. Da GPT-4 bei der Annotation des gesamten Datensatzes einige Äußerungen mit zwei Informationsbedürfnissen kategorisierte, sind manche nicht deutlich einer zuteilbar.

7 Schlussfolgerung und zukünftige Arbeiten

In dieser Arbeit wird eine Methodik entwickelt, welche den Aufwand der Datenannotation reduzieren soll. Mit Hilfe einer Taxonomie-Vorlage aus dem Bereich Kochen wird eine neue für die Domäne DIY erstellt und einige wenige Daten manuell annotiert, um damit durch das Few-Shot Prompting mit GPT-4 die Informationsbedürfnisse in diesem Bereich vorherzusagen und eine große Menge an Annotationsdaten zu erzeugen. Dadurch, dass eine Taxonomie-Vorlage aus einem anderen Themenbereich verwendet und diese an den Kontext angepasst wird, entsteht die zentrale Forschungsfrage „Wie gut lässt sich die Taxonomie von Frummet et al. (2022) auf andere prozedurale Bereiche anwenden?“. Diese Vorlage ist die Ausgangslage, um überhaupt relevante manuelle Annotationsdaten erzeugen zu können, mit denen ein Large Language Model den Annotationsprozess unterstützen und optimieren kann. Durch die Schritte, die durchgeführt werden, um die zentrale Forschungsfrage untersuchen zu können, entstehen zwei weitere Forschungsfragen. Es wird untersucht wie gut mit einer kleinen Menge an annotierten Daten, welche auf einer Taxonomie-Vorlage basieren, die Informationsbedürfnisse im Bereich DIY von einem Large Language Model vorhersagt werden können und ob dies letztlich eine Grundlage darstellt, um mit dem Sprachmodell eine große und relevante Datenmenge erzeugen zu können.
Durch diesen Prozess kann die Generalisierbarkeit der Taxonomie von Frummet et al. (2022) bestätigt werden. Der Ablauf des Prozesses in den beiden verschiedenen Bereichen ist gleich aufgebaut, da sich Nutzer nach den benötigten Arbeitsmitteln und Fakten dazu, sowie dem Ablauf und Durchführung des Prozesses erkundigen. Die Kategorien aus der Koch-Domäne konnten somit fast vollständig übertragen werden, nachdem die domänenspezifischen Begriffe einfach

an die Domäne DIY angepasst wurden. Mit Hilfe dieser neu erzeugten Taxonomie konnte eine repräsentative Stichprobe eines Datensatzes bezüglich der Informationsbedürfnisse manuell annotiert werden. Mit dieser erreichte GPT-4 durch das Few-Shot Prompting einen gesamten *F1-Score* von 93%, wobei es bei der Vorhersage einiger Kategorien Schwierigkeiten gab. Da die Gesamtleistung jedoch sehr überzeugend ist, wurde letztlich der gesamte Datensatz mit dieser Strategie annotiert. Beim Vergleich der beiden Taxonomien wurden bei der Verteilung der Informationsbedürfnisse Unterschiede entdeckt, was darauf hindeutet, dass die beiden Domänen verschiedene Schwerpunkte bezüglich der Informationsbedürfnisse aufweisen. Dies spricht dafür, dass die Taxonomie von Frummet et al. (2022) flexibel und anpassbar an ähnliche Bereiche ist, welche eine gleiche Struktur aufweisen. Dadurch wird die Generalisierbarkeit für prozedurale Bereiche nochmals verdeutlicht.

Einige Limitationen zeigen, dass diese Methodik nochmals optimiert werden könnte, um detailliertere Ergebnisse in der Annotation zu erzielen. Zukünftige Arbeiten könnten realitätsnähere Daten im Bereich DIY sammeln und diese angewandte Methodik nochmals anwenden, um zu untersuchen ob die Ergebnisse von der Methode der Datensammlung abhängen. Ein weiterer prozeduraler Bereich könnte auf die Generalisierbarkeit der Koch-Taxonomie untersucht werden, um eine Generalisierbarkeit der Erkenntnisse dieser Bachelorarbeit zu prüfen.

Literaturverzeichnis

Amatriain, X. (2024). *Prompt Design and Engineering: Introduction and Advanced Methods* (arXiv:2401.14423). arXiv. http://arxiv.org/abs/2401.14423

Balakrishnan, J., & Dwivedi, Y. K. (2024). Conversational commerce: Entering the next stage of AI-powered digital assistants. *Annals of Operations Research*, *333*(2–3), 653–687. https://doi.org/10.1007/s10479-021-04049-5

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4* (arXiv:2312.16171). arXiv. http://arxiv.org/abs/2312.16171

Bu, K., Liu, Y., & Ju, X. (2024). Efficient utilization of pre-trained models: A review of sentiment analysis via prompt learning. *Knowledge-Based Systems*, *283*, 111148. https://doi.org/10.1016/j.knosys.2023.111148

Bunt, H., Petukhova, V., Traum, D., & Alexandersson, J. (2017). Dialogue Act Annotation with the ISO 24617-2 Standard. In D. A. Dahl (Hrsg.), *Multimodal Interaction with W3C Standards* (S. 109–135). Springer International Publishing. https://doi.org/10.1007/978-3-319-42816-1_6

Case, D. O. (2007). *Looking for information: A survey of research on information seeking, needs, and behavior* (2nd ed). Elsevier/Academic Press.

Choi, J. I., Kuzi, S., Vedula, N., Zhao, J., Castellucci, G., Collins, M., Malmasi, S., Rokhlenko, O., & Agichtein, E. (2022). Wizard of Tasks: A Novel Conversational Dataset for Solving Real-World Tasks in Conversational Settings. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Hrsg.), *Proceedings of the 29th International Conference on Computational Linguistics* (S. 3514–3529). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.310

Church, K., & Oliver, N. (2011). Understanding mobile web and mobile search use in today's dynamic mobile landscape. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 67–76. https://doi.org/10.1145/2037373.2037385

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017). SuperAgent: A Customer Service Chatbot for E-commerce Websites. In M. Bansal & H. Ji (Hrsg.), *Proceedings of ACL 2017, System Demonstrations* (S. 97–102). Association for Computational Linguistics. https://aclanthology.org/P17-4017

Devadason, F. J., & Lingam, P. P. (1997). A Methodology for the Identification of Information Needs of Users. *IFLA Journal*, *23*(1), 41–51. https://doi.org/10.1177/034003529702300109

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Hrsg.), *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)* (S. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., & Bing, L. (2023). *Is GPT-3 a Good Data Annotator?* (arXiv:2212.10450). arXiv. http://arxiv.org/abs/2212.10450

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. https://doi.org/10.1186/1472-6947-12-8

Forsythe, D. E., Buchanan, B. G., Osheroff, J. A., & Miller, R. A. (1992). Expanding the concept of medical information: An observational study of physicians' information needs. *Computers and Biomedical Research*, *25*(2), 181–200. https://doi.org/10.1016/0010-4809(92)90020-B

Frummet, A., Elsweiler, D., & Ludwig, B. (2022). "What Can I Cook with these Ingredients?"—Understanding Cooking-Related Information Needs in Conversational Search. *ACM Transactions on Information Systems*, *40*(4), 1–32. https://doi.org/10.1145/3498330

Gao, T., Fisch, A., & Chen, D. (2021). *Making Pre-trained Language Models Better Few-shot Learners* (arXiv:2012.15723). arXiv. http://arxiv.org/abs/2012.15723

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

GovTech Data Science & AI Division. (2023). *PROMPT ENGINEERING PLAYBOOK (Beta v3)*. https://www.developer.tech.gov.sg/products/collections/data-science-and-artificial-intelligence/playbooks/prompt-engineering-playbook-beta-v3.pdf

Gu, J., Zhao, H., Xu, H., Nie, L., Mei, H., & Yin, W. (2023). Robustness of Learning from Task Instructions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Hrsg.), *Findings of the Association for Computational Linguistics: ACL 2023* (S. 13935–13948). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.875

Guy, I. (2016). Searching by Talking: Analysis of Voice Queries on Mobile Web Search. *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 35–44. https://doi.org/10.1145/2911451.2911525

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *Companion Proceedings of the ACM Web Conference 2023*, 294–297. https://doi.org/10.1145/3543873.3587368

Kamvar, M., & Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 701–709. https://doi.org/10.1145/1124772.1124877

Kim, D., & Song, H. (2024). Designing an age-friendly conversational AI agent for mobile banking: The effects of voice modality and lip movement. *International Journal of Human-Computer Studies*, *187*, 103262. https://doi.org/10.1016/j.ijhcs.2024.103262

Labrak, Y., Rouvier, M., & Dufour, R. (2024). A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Hrsg.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (S. 2049–2066). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.185

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations* (arXiv:1909.11942). arXiv. http://arxiv.org/abs/1909.11942

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2021). *A Survey on Text Classification: From Shallow to Deep Learning* (arXiv:2008.00364). arXiv. http://arxiv.org/abs/2008.00364

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). *What Makes Good In-Context Examples for GPT-3?* (arXiv:2101.06804). arXiv. http://arxiv.org/abs/2101.06804

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). *Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity* (arXiv:2104.08786). arXiv. http://arxiv.org/abs/2104.08786

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., & Elazar, Y. (2023). Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Hrsg.), *Findings of the Association for Computational Linguistics: ACL 2023* (S. 12284–12314). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.779

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Oppenlaender, J. (2023). A Taxonomy of Prompt Modifiers for Text-To-Image Generation. *Behaviour & Information Technology*, 1–14. https://doi.org/10.1080/0144929X.2023.2286532

Pecher, B., Srba, I., Bielikova, M., & Vanschoren, J. (2024). *Automatic Combination of Sample Selection Strategies for Few-Shot Learning* (arXiv:2402.03038). arXiv. http://arxiv.org/abs/2402.03038

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation* (arXiv:2102.12092). arXiv. http://arxiv.org/abs/2102.12092

Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3411763.3451760

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications* (arXiv:2402.07927). arXiv. http://arxiv.org/abs/2402.07927

Saka, A. B., Oyedele, L. O., Akanbi, L. A., Ganiyu, S. A., Chan, D. W. M., & Bello, S. A. (2023). Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities. *Advanced Engineering Informatics*, *55*, 101869. https://doi.org/10.1016/j.aei.2022.101869

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *ACM SIGIR Forum*, *33*(1), 6–12. https://doi.org/10.1145/331403.331405

Song, Y., Ma, H., Wang, H., & Wang, K. (2013). Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. *Proceedings of the 22nd International Conference on World Wide Web*, 1201–1212. https://doi.org/10.1145/2488388.2488493

Sun, J., Shaib, C., & Wallace, B. C. (2023). *Evaluating the Zero-shot Robustness of Instruction-tuned Language Models* (arXiv:2306.11270). arXiv. http://arxiv.org/abs/2306.11270

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). *Large Language Models for Data Annotation: A Survey* (arXiv:2402.13446). arXiv. http://arxiv.org/abs/2402.13446

Taylor, R. S. (1962). The process of asking questions. *American Documentation*, *13*(4), 391–396. https://doi.org/10.1002/asi.5090130405

Toney-Wails, A., Schoeberl, C., & Dunham, J. (2024). *AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications* (arXiv:2403.09097). arXiv. http://arxiv.org/abs/2403.09097

Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., Azizi, S., Singhal, K., Cheng, Y., Hou, L., Webson, A., Kulkarni, K., Mahdavi, S. S., Semturs, C., Gottweis, J., … Natarajan, V. (2024). *Towards Conversational Diagnostic AI* (arXiv:2401.05654). arXiv. http://arxiv.org/abs/2401.05654

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). *Want To Reduce Labeling Cost? GPT-3 Can Help* (arXiv:2108.13487). arXiv. http://arxiv.org/abs/2108.13487

Wang, X., Wang, Y., Xu, C., Geng, X., Zhang, B., Tao, C., Rudzicz, F., Mercer, R. E., & Jiang, D. (2023). *Investigating the Learning Behaviour of In-context Learning: A Comparison with Supervised Learning* (arXiv:2307.15411). arXiv. http://arxiv.org/abs/2307.15411

Wang, Y., & Luo, Z. (2023). Enhance Multi-Domain Sentiment Analysis of Review Texts Through Prompting Strategies. *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, 1–7. https://doi.org/10.1109/HDIS60872.2023.10499502

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., & Young, S. (2017). *A Network-based End-to-End Trainable Task-oriented Dialogue System* (arXiv:1604.04562). arXiv. https://doi.org/10.48550/arXiv.1604.04562

Wilson, T. D. (1981). On User Studies and Information Needs. *Journal of Documentation*, *37*(1), 3–15. https://doi.org/10.1108/eb026702

Yamada, M. (2023). Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In M. Yamada & F. do Carmo (Hrsg.), *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track* (S. 195–204). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2023.mtsummit-users.19

Yi, J., & Maghoul, F. (2011). Mobile search pattern evolution: The trend and the impact of voice queries. *Proceedings of the 20th International Conference Companion on World Wide Web*, 165–166. https://doi.org/10.1145/1963192.1963276

Young, I. J. B., Luz, S., & Lone, N. (2019). A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*, *132*, 103971. https://doi.org/10.1016/j.ijmedinf.2019.103971

Zhang, P., & Boulos, M. N. K. (2023). Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. *Future Internet*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. *Proceedings of the 38th International Conference on Machine Learning*, 12697–12706. https://proceedings.mlr.press/v139/zhao21c.html