Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Verwendung und Auswertung von generativer KI zur Generierung von Erklärungen für Software Systeme

Using and Evaluating Generative AI to Generate Explanations for Software Systems

Masterarbeit

im Studiengang Informatik

von

Pia Brandt

Prüfer: Prof. Dr. Kurt Schneider
Zweitprüfer: Dr. Jil Klünder
Betreuer: Hannah Deters, M.Sc.

Hannover, 22.04.2024

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 22.04.2024

Pia Brandt

Zusammenfassung

Künstliche Intelligenz wird immer wichtiger und mit den neuesten generativen Modellen wie ChatGPT einfacher für die Allgemeinheit zugänglich. Auch unabhängig davon werden bisherige und neue Software Systeme immer umfangreicher und komplexer. Die Hilfestellung für den Anwender bleibt hierbei häufig auf der Strecke, umfangreiche und abdeckende Erklärungen zu entwickeln wird oft aus kosten- und Zeitgründen vernachlässigt. Anbindung von so genannten Large Language Models können als Hilfestellung dienen, um Antworten für Erklärungsbedarf von Nutzern zu generieren.

Zur Untersuchung der Güte der Erklärungen wurden zunächst auf einem Qualitätsmodell basierende wichtige Aspekte erarbeitet und Metriken aufgestellt. In einer Vorstudie von Droste et al. wurde Erklärungsbedarf erhoben. Darauf aufbauend wurden Anfragedaten entwickelt und Erklärungen generiert. Mit diesen Anfragen und Erklärungen erfolgte eine Online-Nutzerstudie, deren Ergebnisse analysiert und ausgewertet wurden. Zur Ergänzung dieser Ergebnisse wurden zusätzliche Interviews im gewohnten Arbeitsumfeld bei ausgewählten „geübten Vielanwendern“ durchgeführt. In diesen Interviews konnten die Anwender Fragen an ChatGPT stellen, welche sich auf eine von ihnen ausgewählte Software sowie vordefinierte Aufgabe bezogen. Insgesamt wird die Nutzung von generativer Künstlicher Intelligenz zur Unterstützung der Anwender durch Erklärungen als sehr zielführend erachtet. Die Erklärungen von ChatGPT wurden als zufriedenstellend wahrgenommen, die Effektivität und der Wunsch nach Software-Integration fiel hingegen unterschiedlich aus. Der hier untersuchte Ansatz bietet aufgrund des selbstlernenden Charakters und dadurch höhere Aktualität potentiell auch Vorteile und Arbeitsersparnis auf Seiten der Softwareentwicklung.

Abstract

Artificial Intelligence is becoming increasingly important and, with the latest generative models such as ChatGPT, more accessible to the general public. Regardless of these developments, existing and new software systems are becoming more extensive and complex. Support for the user often seems not to be sufficent, and the development of extensive and comprehensive explanations is often neglected for reasons of cost and time. Integrating so-called Large Language Models serves as a tool to generate answers to users' explainability needs.

Important aspects based on a quality model were first developed and metrics established to investigate the quality of the explanations. A preliminary study by Droste et al. identified the explainability needs. Inquiry data was developed and explanations generated on the basis of the explainability needs identified in a preliminary study. Based on this, query data was developed and explanations generated. These queries and explanations were used to conduct an online user study, the results of which were analyzed and evaluated. To supplement these results, additional interviews were conducted in the usual working environment with selected „power users“. In these interviews, the users were able to ask ChatGPT questions relating to a software they had selected and a predefined task. Overall, the use of generative Artificial Intelligence for support users with explanations is considered to be very effective. The explanations provided by ChatGPT were perceived as satisfactory, while the effectiveness and the desire for software integration varied. The approach examined here also offers potential advantages and labor savings on the software development side due to its self-learning character and thus higher actuality.

Inhaltsverzeichnis

1 Einleitung **1**
- 1.1 Problemstellung . . . 1
- 1.2 Lösungsansatz . . . 1
- 1.3 Struktur der Arbeit . . . 2

2 Grundlagen und Verwandte Arbeiten **3**
- 2.1 Erklärbarkeit . . . 3
- 2.2 Erklärungen . . . 4
- 2.3 Metriken zur Berechnung von Verständlichkeit . . . 5
- 2.4 Generative KI . . . 6
- 2.5 Verwandte Arbeiten . . . 7

3 Metriken zur Erklärbarkeit **13**
- 3.1 Zieldefinition und Methodik . . . 13
  - 3.1.1 Anwendungsszenario und Nutzer . . . 13
  - 3.1.2 Forschungsfragen . . . 15
  - 3.1.3 Vorgehensweise . . . 17
- 3.2 Metriken . . . 17
  - 3.2.1 Aspekte der Erklärbarkeit . . . 18
  - 3.2.2 Kriterien und Metriken der Aspekte . . . 19
  - 3.2.3 Ideen anderer Metriken . . . 24
- 3.3 Fragen für die Online-Studie . . . 26

4 Entwicklung der Umfrageinhalte **29**
- 4.1 Promptentwicklung . . . 29
- 4.2 Antwortengenerierung . . . 34
- 4.3 Vorbereitungen zur Studie . . . 36

5 Nutzerstudie I: Online-Studie **37**
- 5.1 Methodik der Online-Studie . . . 37
  - 5.1.1 Entwurf . . . 38
  - 5.1.2 Implementierung . . . 39
  - 5.1.3 Durchführung . . . 39
  - 5.1.4 Datenanalyse . . . 41

5.2 Ergebnisse der Online-Studie 44
5.2.1 Bewertung der Erklärungen: Frage-Antwort-Paare . . 44
5.2.2 Gesamtbewertung: Abschlussseite 50

6 Nutzerstudie II: Interview-Studie 53
6.1 Methodik der Interview-Studie 53
6.1.1 Zieldefiniton 54
6.1.2 Teilnehmerakquise 55
6.1.3 Erstellung des Leitfadens 56
6.1.4 Durchführung 60
6.1.5 Analyse 61
6.2 Ergebnisse Interview-Studie 62

7 Sonstige Auswertungen zu Metriken 65
7.1 Lesbarkeitsindizes 65
7.2 Manuelle Auswertung verbliebener Metriken 68

8 Diskussion 69
8.1 Beantwortung der Forschungsfragen 69
8.2 Betrachtung der Hypothesen 72
8.3 Limitationen 74

9 Fazit 77
9.1 Zusammenfassung 77
9.2 Ausblick 78

A Online-Studie 81
A.1 Screenshots der Online-Studie 81
A.2 Software Systeme 85

B Interview-Studie 87
B.1 Leitfaden für den Teilnehmer 87
B.2 Leitfaden der Interview-Studie 90

C Weitere Grafiken zur Auswertung 97

Kapitel 1

Einleitung

Die Entwicklung von Künstlicher Intelligenz (KI) hat in den letzten Jahren einen rasanten Anstieg genommen. Mithilfe von auf Large Language Modelle (LLM) basierenden Ansätzen können natürlichsprachliche Texte generiert werden. Diese fallen unter den Bereich der generativen KI. Für Aufsehen sorgte OpenAI in der allgemeinen Öffentlichkeit durch Veröffentlichung von ChatGPT. Der Chatbot wurde dazu entwickelt, Inhalte zu generieren [10], als wären sie Menschen erzeugt worden und zeigte dabei erstaunliche Ergebnisse [39].

1.1 Problemstellung

Unabhängig der voranschreitenden Entwicklung von KI nimmt die Komplexität von Software Systemen in der heutigen Zeit weiter zu. Die Erklärbarkeit eines Systems hat Einfluss auf die Transparenz des Systems und damit auch auf das damit verbundene Vertrauen des Nutzers [8, 7]. Der beim Anwender zunehmend entstehende Erklärungsbedarf [16], kann kaum gedeckt werden. Entwickler müssten bereits in der Anforderungsanalyse diverse Variablen zur Erklärbarkeit bedenken [8], was mit zunehmenden Entwicklungsaufwand und -kosten verbunden ist. Doch um ein System erklärbar zu gestalten, müssten Erklärungen für den Nutzer zugänglich gemacht werden [30]. Abhilfe könnte hier generative KI schaffen, deren Eignung es zu prüfen gilt.

1.2 Lösungsansatz

In dieser Arbeit wird untersucht, inwieweit Erklärungsbedarf bzw. Anwendungsprobleme bei dem täglichen Gebrauch von Software Systemen statt von festgelegten, statischen Erklärungen durch eine Anbindung von generativer KI dynamisch geklärt werden können. Ein großer Vorteil dieses Ansatzes besteht darin, dass Anforderungen an die Erklärbarkeit sowie die Erklärungen an sich nicht zuvor entwickelt werden müssten, sondern

durch Nutzerinteraktion angepasst an den Nutzer, den Kontext und das Ziel generiert würden. Aufgrund des selbstlernenden Charakters wissensbasierter Systeme können die gelieferten Erklärungen stets aktuell gehalten und kontinuierlich verbessert werden, ohne dass laufend seitens des Softwareherstellers manuelle Aktualisierungen der Erklärungen selbst erfolgen.

Hierbei wird in dieser Arbeit insbesondere die Güte der generierten Erklärungen untersucht, und nicht die Umsetzung als Software Integration als solches. Dabei wird neben der Zufriedenheit der Nutzer mit den Erklärungen sowie deren Anwendbarkeit auch das Vertrauen des Nutzers in die generierende KI ChatGPT betrachtet.

1.3 Struktur der Arbeit

Die Arbeit ist wie folgt strukturiert: Zunächst werden Grundlagen zu Erklärbarkeit und Erklärungen sowie Generativer KI erläutert. Anschließend wird die Methodik zur Beantwortung der dort aufgestellten Forschungsfragen vorgestellt. Hier werden wichtige Aspekte zur Bewertung von Erklärbarkeit anhand eines Qualitätsmodells identifiziert und davon ausgehend Metriken erstellt. In Kapitel 4 wird die Entwicklung von Anfragen und Prompts beschrieben, welche für die weiteren Untersuchungen verwendet wurden. Die Entwicklung und Durchführung der Online-Studie sowie deren Ergebnisse werden dargestellt. Weiter werden Methodik und Ergebnisse der vertiefenden Interview-Studie erläutert. Die Bewertung verbliebener Metriken wird vorgestellt. In der Diskussion der Ergebnisse werden die Forschungsfragen beantwortet, aufgestellte Hypothesen untersucht und auf die Limitationen der Arbeit eingegangen. Abschließend folgt eine Zusammenfassung und es wird ein Ausblick auf weitere Forschungsansätze gegeben, welche auf der Verwendung von generativer KI basierten.

Kapitel 2

Grundlagen und Verwandte Arbeiten

In diesem Kapitel werden grundlegende Begriffe zur Erklärbarkeit, Erklärungen und verwendeter Metriken festgehalten. Außerdem wird definiert, was eine generative KI ist und der für diese Arbeit verwendete Chatbot ChatGPT wird vorgestellt. Im Anschluss wird ein Blick auf verwandte Arbeiten geworfen, welche sich mit ChatGPT, Erklärungen und Erklärbarkeit befasst haben.

2.1 Erklärbarkeit

Erklärbarkeit (engl. Explainability) ist die Fähigkeit einer Software, Erklärungen zu liefern bzw. für einen Adressaten erklärbar zu sein [9, 13]. Ziel von Erklärbarkeit ist also, dem Menschen das Verständnis für verschiedene Aspekte von software-gesteuerten Systemen zu ermöglichen [30]. Als nicht-funktionale Anforderung (NFR) [30] ist Erklärbarkeit, wie andere NFRs, schwer zu erheben [8, 30]. Für das dafür zunächst benötigte Verständnis definiert Chazette et al. erklärbare Systeme wie folgt:

Definition Erklärbarkeit [8]

A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C.

Während das System S, zu welchem Erklärungen benötigt werden, in der Regel feststeht, können der zu erklärende Aspekt X, der Adressat A, der Kontext C und die Entität E unterschiedlichste Werte annehmen [8]. Entsprechend kann ein System in bestimmten Hinsichten erklärbar sein und in anderen nicht [30]. Die richtige Art von Erklärbarkeit muss in der Anforderungsanalyse herausgefunden und sich der unterschiedlichen Werte der Variable bewusst gemacht werden [8]. Die Umsetzung der Erklärbarkeit beeinflusst signifikant viele andere Qualitätsaspekte [8, 9, 30] und kann hierbei als „Mittel zum Zweck“ für die Erreichung anderer Qualitätsaspekte von Software Systemen gesehen werden [13]. Positiv beeinflusst werden unter anderem die Transparenz, Akzeptanz und Verständlichkeit des Systems sowie die Genauigkeit des mentalen Modells, die Entscheidungsfindung und das Vertrauen des Nutzers in das System [8]. Dagegen können Erklärbarkeitsanforderungen mit Entwicklungskosten, Präzision und Performance in Konflikt stehen [30]. Wird beispielsweise der Fokus darauf gelegt, jede Funktion einer Software zu erklären, so steigert dies die Transparenz der Software. Dadurch wird möglicherweise auch das Verständnis und Vertrauen des Nutzers dem System gegenüber gesteigert [7]. Gleichzeitig entsteht ein hoher Entwicklungsaufwand und damit verbundener erhöhter Ressourcenaufwand.

2.2 Erklärungen

Zur Bewertung der Erklärbarkeit sollte sich von vornherein Klarheit darüber verschafft werden, welche „Gebrauchsweisen von Erklärungen analysiert und präzisiert“ werden sollen, so Stegmüller [41]. Eine allgemeine Begriffsdefinition sei sehr allgemein und unbestimmt: Erklärungsbedarf besteht, wenn „der Fragende in irgendeiner Weise verwirrt, verblüfft, ‚konsterniert‘ ist“; eine Erklärung dient dazu, „seine Verwirrungen zu beseitigen“ [41]. Aus dieser Charakterisierung ist zu entnehmen, dass Erklärungen sich nach dem Erklärungsbedarf richten und auch dieser entsprechend unterteilt werden kann.

Für die aktuelle Arbeit ist eine grobe Unterscheidung in drei Erklärungstypen ausreichend, welche unter anderem auch Klein [28] und Debelak [11] verwenden:

- **Warum-Erklärungen** dienen der *kausalen Erklärung von Vorgängen oder Tatsachen* [41]. Der Fragende will Zusammenhänge verstehen oder Ursachen und Konsequenzen nachvollziehen. Der Nutzer könnte zum Beispiel fragen „warum das Laden von Inhalten trotz guter Internetverbindung so lange dauert“. Eine Erklärung sollte dann auf mögliche Gründe eingehen bzw. zu einem Verständnis im engeren Sinne führen [11].

- **Was-Erklärungen** sind zur *Erklärung der Bedeutung eines Wortes* oder *Klarlegen des Sinnes* in Form einer *ungefähren Erläuterung* oder einer *Definition* im wissenschaftlichen Sinn [41]. Eine Beispielfrage des Nutzers könnte sein: „Was sind Pivot-Tabellen?“. Die Erklärung soll zu einer Erkenntnis führen und den Fragenden befähigen, Zusammenhänge erkennen zu können [11].

- **Wie-Erklärungen** enthalten eine *mehr oder weniger detaillierte Schilderung* zu beispielsweise einer Handlung oder Funktion [41]. Beispielhaft könnte eine Frage des Nutzers sein: „Wie werden Pivot-Tabellen erstellt?“. Die Erklärung soll zur Handlungsfähigkeit führen und der Instruktion und Anleitung dienen [11].

2.3 Metriken zur Berechnung von Verständlichkeit

Zur Bewertung der Verständlichkeit eines Textes, und somit auch einer Erklärung, können verschiedene Lesbarkeitsindizes zu Rate gezogen werden. Mittels einer Formel wird anhand verschiedener Anhaltspunkte, wie beispielsweise Anzahl Wörter und Länge des Textes, ein Wert bzw. Index berechnet. Das Ergebnis gibt dabei an, wie schwierig ein Text zu lesen ist. Bekannte Metriken für die englische Sprache sind der Flesch Reading Ease Score, das Flesch-Kincaid Grade Level [27] und der Gunning Fog Index [21]. In der Arbeit von Briest [3] wurden diese und weitere Indizes zur Lesbarkeit verglichen und das Korrelationswichtungsmaß als weiterer Ansatz entwickelt.

Flesch Reading Ease (FRE) Score [18]:

Der Flesch Reading Ease Score wurde zur Berechnung der Lesbarkeit eines englischen Textes entwickelt. Für die deutsche Sprache veröffentlichte Amstad [1] eine angepasste Variante:

$$FRE_{de} = 180 - 1,0 * \frac{\#W\ddot{o}rter}{\#S\ddot{a}tze} - 58,5 * \frac{\#Silben}{\#W\ddot{o}rter}$$

Der FRE Score spiegelt wieder, wie schwer ein Text zu lesen ist und bezieht dabei insbesondere die Satzlänge und Silbenanzahl in die Berechnung mit ein. Die Werte werden anhand verschiedener Zahlenbereiche, zwischen 0 und 100 liegend, in Schwierigkeitsgrade eingeteilt. Negative Werte werden dabei auf 0 gesetzt. Ein Text mit einem Wert bis 30 wird als *sehr schwer* lesbar und entsprechend von akademischem Niveau eingeordnet. Ein Wert zwischen 60 und 70 entspricht einer *mittleren* Lesbarkeit, welche jedoch als gut verständlich ab einem Niveau der 10. Klasse angesehen wird.

Korrelationswichtungsmaß (KWM) nach Briest [3]:

Das KWM bezieht im Vergleich zu anderen genannten Formeln zusätzliche Variablen in die Berechnung eines Lesbarkeitswertes mit ein. Neben der Satzlänge wird auch die Anzahl Satzglieder (wie Subjekt, Prädikat, Objekt, adverbiale Bestimmungen) in der Lesbarkeitsberechnung hoch gewichtet. Außerdem mit einbezogen werden die Anzahl *Fremdwörter*, *Abstrakta*, *substantivische Attribute* und die *Länge des Satzrahmens*, welche negativen Einfluss auf die Lesbarkeit haben. Als einzig positiv zur Verständlichkeit betragend wird in der Metrik die Anzahl Verben bzw. die Verbintensität pro Satz betrachtet.

Flesch Reading Ease Score und KWM wiesen in den Untersuchungen von Briest [3] eine geringe Korrelation auf. In einem dort durchgeführten Versuch sollten 200 Rundfunkmitarbeiter insgesamt 160 Sätze auf ihr Verständlichkeit hin einschätzen. Das KWM wies zu den gemittelten Schätzurteilen eine Korrelation von 0,86 auf.

Weitere Metriken zur Berechnung der Verständlichkeit werden aufgrund ähnlichen Aussagewertes nicht näher betrachtet.

2.4 Generative KI

Lim et al. [32] definieren generative KI als „Technologie, die (i) Deep-Learning-Modelle nutzt, um (ii) menschenähnliche Inhalte (z.B. Bilder, Wörter) als Antwort auf (iii) komplexe und vielfältige Anfragen [bzw. Prompts] (z.B. Sprachen, Anweisungen, Fragen) zu erzeugen“. Dabei werden große Mengen existierender (Trainings-)Daten auf Muster sowie Verteilungen analysiert und diese mittels Deep-Learning Verfahren gelernt [2, 5, 17]. Ein wichtiger Teil generativer KI sind Generative Pre-trained Transformer (GPT) Modelle [2].

GPTs werden durch digital verfügbare Text-Daten vortrainiert und erzeugen qualitativ hochwertige Ausgaben, welche auf spezielle Zwecke anpassbar sind [44]. Sie werden vor allem für Aufgaben zur Verarbeitung von natürlicher Sprache (Natural Language Processing, NLP) verwendet [2, 37, 44]. Neben dem Generieren und Klassifizieren von Texten kommen sie auch zur menschenähnlichen Konversation [2], Klärung von Fragen und dem Schreiben kreativer [44] oder formaler Inhalte [38] zum Einsatz.

Openai.org veröffentlichte im November 2022 den Chatbot ChatGPT, welcher zunächst auf GPT3 basierte [2, 4, 17]. Dieser baut auf Large Language Models (LLM) auf [17, 39], welche mittels eines mathematischen Modells das nächste Wort vorhersagen. Der Prompt setzt dabei einen initialen Wert [17]. Es können neue Inhalte basierend auf dem gelernten Wissen und Kontext generiert werden [39] und bemerkenswerte Leistungen auch ohne Aufgaben-spezifisches Training erzielt werden [25]. Kaylan [25]

gibt eine Übersicht über die Entwicklung der GPT-3 family large language models (GLLMs), inklusive ChatGPT und der aktuelleren Version GPT-4.

Stand April 2024 verwendete ChatGPT in der kostenfreien Version GPT-3.5; ein kostenpflichtiges Upgrade zu Version 4 war möglich. Die Anwendungsmöglichkeiten von ChatGPT sind divers und werden im folgenden Abschnitt

der Verwandten Arbeiten näher betrachtet. Wie mehrere Arbeiten beschreiben, neigt ChatGPT zum Halluzinieren [20, 36]. Damit ist gemeint, dass Informationen in Zusammenhang gebracht werden, welche für sich genommen korrekt sein können, jedoch in der Kombination falsch sind. Ein Beispiel hierfür sind Literaturangaben, bei denen Autoren nicht zu den Veröffentlichungen passen oder Zweitautoren als Erstautoren aufgeführt werden [36].

2.5 Verwandte Arbeiten

In der Literatur werden die sich ergebende Möglichkeiten und Herausforderungen bei der Verwendung von ChatGPT in verschiedenen Anwendungsgebieten untersucht.

Für die Verwendung zur Generierung von Erklärungen zu Software Systemen scheint es bisher wenig vergleichbare Veröffentlichungen zu geben. In dieser Arbeit wird KI daraufhin untersucht, Systeme erklärbarer zu gestalten. Das ist nicht zu verwechseln mit explainable AI (XAI), welche versucht, KI selbst zu erklären [29].

Potenzial von ChatGPT zur Generierung von Erklärungen

Die Arbeit von Mavrepis et al. [34] beschäftigt sich damit, XAI für aller Art Nutzer zugänglicher zu machen. Dazu wird der ChatGPT Builder genutzt, um ein selbst definiertes Large Language Model zu erstellen und damit bessere Ergebnisse zu erzielen. Dieses soll angepasst an das Wissen und Interesse des Nutzers verschiedene XAI Methoden klar und prägnant erklären bzw. zusammenfassen. Zur Ermittlung der Anwendbarkeit und Effektivität des Modells wurde eine Studie mit möglichst breitem Spektrum von Fachleuten durchgeführt. Das Ergebnis war, dass das erstellte Modell effektiv darin sei, leicht verständliche, zielgruppenangepasste Erklärungen zu liefern.

Naumann et al. [36] untersuchten in ihrer Arbeit die Qualität bzw. Zuverlässigkeit und das Unterstützungspotenzial von ChatGPT im Hinblick auf nachhaltige Software-Entwicklung. Dabei wurden die Antworten hinsichtlich der Kriterien Verständlichkeit, Korrektheit, Vollständigkeit und Verwendbarkeit von den Autoren bewertet. Insgesamt waren die Antworten grundsätzlich schlüssig und richtig oder wiesen zumindest in die richtige Richtung. Gleichzeitig waren sie häufig oberflächlich oder unkonkret mit

besonderer Schwäche im Halluzinieren von Quellen. Zum Beispiel wurden Kombinationen von Autoren und Publikationen aufgeführt, die so nicht zusammen gehörten oder existierten. Naumann et al. kamen zu dem Schluss, dass ChatGPT für einen Einstieg und Überblick über das untersuchte Thema durchaus nutzbar sei.

Ähnliche Ergebnisse beobachtete Geist [20] in seiner Betrachtung von ChatGPT im Bereich der Rechtsinformation. ChatGPT lerne zwar, könne jedoch nicht zwischen richtig und falsch differenzieren. Gefährlich hierbei ist, dass Aussagen logisch erscheinen, aber dennoch falsch sein können. Trotz fehlender Referenzen und Halluzinierens kann ChatGPT als unterstützendes Werkzeug zum Einsatz kommen. Ein sinnvoller, aber kritischer Umgang ist anzustreben, so sein Schluss.

In einer anderen Arbeit von Brynjolfsson et al. [5] wurde der Einfluss auf die Arbeitsleistung von unerfahrenen und erfahreneren Arbeitskräften untersucht. Hier könne ChatGPT vor allem die Produktivität von Neueinsteigern und weniger Qualifizierten steigern. Für Experten ergibt sich geringere Hilfe mit minimalen Effekten. Die Kundenstimmung und der Lernprozess der Angestellten seien außerdem positiv beeinflusst.

Kabir et al. [24] betrachteten den Einsatz von ChatGPT insbesondere im Bezug auf Programmierfragen. Die Autoren untersuchen die Antworten von ChatGPT auf Stack Overflow Fragen hinsichtlich der Kriterien Korrektheit, Konsistenz, Verständlichkeit und Prägnanz durch eine Nutzerstudie sowie eine linguistische Analyse. Insgesamt enthielten hier mehr als die Hälfte der Antworten inkorrekte Informationen, darunter hauptsächlich konzeptionelle Fehler. Antworten waren zumeist ausschweifend, aber gut verständlich. Dabei schien ChatGPT sich formal, analytisch und in positiver Stimmung auszudrücken. In der Nutzerstudie schnitt Stack Overflow in den untersuchten Kriterien besser ab als ChatGPT. Stack Overflow wurde von den Teilnehmern mehrheitlich bevorzugt. Dennoch präferierten manche Nutzer die Antworten von ChatGPT, darunter auch falsche, aber überzeugende Antworten.

Dongmo et al. [14] untersuchten den Einsatz von ChatGPT im Hochschulalltag. Als mögliche Einsatzgebiete fassen sie die Erzeugung von Texten, Ideengenerierung, Verwendung als Lernhilfe bzw. Tutor und Unterstützung beim Programmieren zusammen. Im Gegensatz zu anderen Arbeiten sehen sie den Zugang zu KI als Chance, die eigene digitale Kompetenz zu verbessern und bewusst mit Risiken umzugehen. Es werden unter anderem die mangelnde Integration von wissenschaftlichen Quellen, die Verbreitung von Falschinformationen und eine mögliche missbräuchliche Nutzung als problematisch angesehen. Außerdem sei die Möglichkeit der Autorenschaft von ChatGPT generierten Texten unklar. Insgesamt wird ChatGPT als hilfreiche Ressource angesehen, welche jedoch keine inhaltliche Prüfung vornimmt, sondern vielmehr Sprache analysiert und produziert.

In der Arbeit von Michel-Villareal et al. [35] werden ebenfalls im Bildungsbereich Chancen und Herausforderungen von ChatGPT betrachtet. Insbesondere fragliche Integrität, Plagiat-Erkennung und ein möglicher Einfluss auf das eigene kritische Denken werden kritisch zum Ausdruck gebracht. Es besteht das Risiko, sich zu sehr auf ChatGPT zu verlassen, welcher falsche Informationen verbreiten kann. Genauigkeit und Verlässlichkeit werden hierbei als Schlüsselprobleme genannt. Neben möglichem Bias im Modelltraining ist auch aufgrund des Vortrainierens eine Aktualität der Wissensbasis nicht gegeben. Es wurde ein semi-strukturiertes Interview mit ChatGPT als Befragten durchgeführt, auf dessen Basis die Integration von ChatGPT in den Hochschulalltag bewertet wurde. ChatGPT kann Unterstützung sowohl für die Studenten als auch für die Lehrenden bieten. Die oben genannten Bedenken wurden durch die von Michel-Villareal et al. durchgeführten Studie als Herausforderungen bestätigt.

Fuchs [19] dokumentiert ähnliche Ergebnisse. Auch er sieht Potenzial darin, Studenten personalisiert zu helfen und beim Lernen zu unterstützen. Gleichzeitig gilt es ein ‚Übervertrauen' und Sinken des kritischen Denkens zu berücksichtigen. Fuchs sieht außerdem ebenfalls die Genauigkeit von ChatGPT als Herausforderung; ChatGPT hat Probleme mit Nuancen der Komplexität menschlicher Sprache, wodurch es zu Missverständnissen und falschen Antworten kommt.

Eine zusammenfassende Übersicht potenzieller Vor- und Nachteile zum Einsatz von ChatGPT beim Lehren und Lernen geben Baidoo-Anu und Ansah [2] sowie Kasneci et al [26]. Lim et al. [32] betiteln Teile davon auch als paradox. Nach diesen ist ChatGPT bzw. generative KI „ein Freund und doch ein Feind" [32]. Hierunter fällt, dass KI als Tool das Schreiben und die Wissensabfrage erleichtert und eine sehr gute Unterstützung bieten kann. Auf der anderen Seite sei es schwer, generierte von menschlichen Inhalten zu unterscheiden. Lim et al. [32] merken an, dass das eigene Denken womöglich vernachlässigt würde und generierte Inhalte bis dato sehr fehleranfällig seien. Weiter heißt es, generative KI sei „fähig und doch abhängig"[32]. Während es leicht fällt, Texte zu analysieren, zu strukturieren und darzustellen, sind die gegebenen Antworten dahingegen von der Eingabe bzw. dem Prompt abhängig. Neben allgemeiner Fehleranfälligkeit spielen auch Inkohärenz und Prägnanz eine Rolle. Weiterhin werden das Verbannen bzw. Verbieten von generativer KI entgegen der rasch zunehmenden Popularität sowie der Zugang zu der Technologie abgehandelt. Die Meinung über den Einsatz ist generell kontrovers; auch andere Autoren geben die Verantwortlichkeit, Integrität und ethische Aspekte zu bedenken [33, 38]. Das kritische Denken, Interpretieren und Diskutieren eines Wissenschaftlers kann und sollte ChatGPT nicht übernehmen, so Quintans-Júnior et al. [38]. Hier gibt es widersprechende Literatur; Lim et al. [32] sehen generative KI als Chance und eine Zukunft als *„game-changer"*, auch im wissenschaftlichen Kontext bzw. Bildungsbereich und der Co-Autorenschaft.

Bewertung von Erklärungen bzw. Erklärbarkeit, Rolle der Erklärbarkeit

Krech [29] geht in seiner Arbeit auf den vermeintlichen Konflikt zwischen der Erklärbarkeit und der Performance von KI Modellen ein. Für einen reflektierten Einsatz und verantwortungsvollen Umgang mit Daten und der KI selbst seien zentrale Grundpfeiler zu beachten. Dazu gehört neben Transparenz, Reproduzierbarkeit und Fairness auch Erklärbarkeit, welche sich gegenseitig bedingen. Erklärbarkeit trage zur Verbesserung der Modelle bei. Außerdem würde das Vertrauen und die Akzeptanz beim Anwenders gesteigert. Erklärbarkeit trage beim Verhindern von Bias und Vorurteilen bei.

Speith et al. [40] gehen auf eine genauere Evaluierung von Erklärbarkeitsansätzen ein. Hier wird zunächst eine Klassifizierung der Evaluierungsmethoden (EMs) nach Doshi-Velez und Kim [15] vorgenommen. Anwendungsbezogene EMs evaluieren dabei den Einfluss auf Experten in bestimmten Aufgaben. Menschbezogene EMs bewerten die Auswirkung auf einen beliebigen Menschen in einer generellen Umgebung. Diese beiden Ansätze werden durch Nutzerstudien bewertet. Die Evaluierung dieser EMs ist zeit- und kostenintensiver, geben aber Einsicht in die eigentlichen Auswirkungen der Erklärungen. Funktionalbezogene EMs nutzen mathematische Vorgaben und Tests bzw. Heuristiken. Diese vermeiden Probleme, wie sie beim Einbezug von Nutzern entstehen, wie zum Beispiel Nutzer Bias, mangelnde Generalisier- und Replizierbarkeit sowie Limitationen beim Studiendesign. Da Erklärbarkeit jedoch dazu gedacht ist, Menschen zu helfen, sollten die nutzerzentrierten Ansätze nicht außer Acht gelassen werden. Speith et al. [40] befinden Nutzerzufriedenheit und -verständnis der erhaltenen Informationen als wichtige zu erfassende Aspekte. Abschließend seien die besten Evaluierungsmethoden abhängig vom Kontext und sollten aus einer Kombination von verschiedenen EMs bestehen.

Lee at al. [31] versuchten sich an einer objektiven Bewertung der Qualität von ChatGPTs Antworten durch eine Expertenanalyse. Dabei erhielten vier Mediziner, erfahrene und weniger erfahrene, zufällig Antworten von Krankenhauswebsites oder ChatGPT zu häufig gestellten Fragen ihres Fachgebiets. Diese sollten sie auf Verständlichkeit, wissenschaftliche Angemessenheit und ihre Zufriedenheit hin bewerten. Verwendet wurden dafür 7-stufige Likert Skalen. In allen Bereichen schnitten KI und nicht-KI Antworten ähnlich ab, mit nicht signifikant höherer Qualität der KI Antworten. Einem Teilnehmer fiel es leicht, generierte Antworten zu erkennen. Die anderen konnten diese nicht von den Antworten der Websites unterscheiden. Die Studie eigt das Potenzial der Verwendung von beispielsweise ChatGPT zur Optimierung der Kommunikation, hier zwischen Patient und Gesundheitswesen. Außerdem wurde das Lesbarkeitslevel der Antworten mittels der Flesch-Kincaid und Gunning Fog Metriken bestimmt. Dabei stellte sich heraus, dass die

generierten Antworten einen signifikant höheren Lesbarkeitsindex aufwiesen und somit schwerer zu lesen waren.

Deters et al. [12] entwarfen ein nutzerzentriertes Qualitätsmodell zur Bewertung von Erklärbarkeit, auf welchem diese Arbeit insbesondere aufbaut. Für das Qualitätsmodell werden zehn Aspekte der Erklärbarkeit identifiziert und entsprechende Kriterien und Metriken zur Analyse aufgestellt. Die Aspekte sind Verständlichkeit, Transparenz, Effektivität, Effizienz, Zufriedenheit, Korrektheit, Angemessenheit, Vertrauenswürdigkeit, Überzeugungskraft und Überprüfbarkeit. Verständlichkeit bezieht sich darauf, ob die Erklärungen vom Adressaten einfach zu verstehen sind. Transparenz, ob die Erklärungen ausreichend Einblick in die Funktionsweise des Systems gibt. Effektivität beschreibt, inwieweit die Erklärung den Adressaten befähigt, das System besser zu nutzen und entsprechende Entscheidungen zu treffen. Bei der Effizienz geht es um die schnellere Nutzung. Die Zufriedenheit spiegelt wider, ob die Erklärung Nutzungskomfort und Vergnügen steigert. Angemessen ist eine Erklärung, wenn sie Kontext, Nutzer sowie Nutzungsziel entsprechend passend adressiert. Eine Erklärung kann dabei helfen, dass das Vertrauen des Adressaten in das System steigt. Überzeugungskräftige Erklärungen überzeugen den Adressaten, etwas Systembezogenes zu tun. Die Überprüfbarkeit ermöglicht dem Nutzer ein Eingreifen in das System sowie das Korrigieren beim Auftreten von Fehlern. Da das Modell nutzerzentriert ist, greifen viele der Metriken auf die Beteiligung des Nutzers zurück. Vor allem bei der Zufriedenheit und dem Vertrauen ist es wichtig, die Nutzerwahrnehmung zu erfassen.

Kapitel 3

Metriken zur Erklärbarkeit

Wie in den verwandten Arbeiten in Abschnitt 2.5 beschrieben, gibt es diverse Ansätze, Erklärbarkeit zu messen. Jedoch wurde bis zum aktuellen Zeitpunkt kein einheitliches Verfahren entwickelt. Für diese Arbeit wird sich insbesondere auf das Qualitätsmodell von Deters et al. [12] bezogen und davon ausgehend wichtige Aspekte und Metriken abgeleitet. Dazu wird zunächst das übergeordnete Ziel und die Methodik festgehalten.

3.1 Zieldefinition und Methodik

Da die Erklärungen vom Nutzer, dem Kontext [8] sowie dem angestrebten Ziel [13] abhängig sind, werden diese näher definiert. Außerdem wird ein Überblick darüber gegeben, wie das Vorgehen zur Klärung der Forschungsfragen in dieser Arbeit ist.

3.1.1 Anwendungsszenario und Nutzer

Das gedachte Anwendungsszenario sieht eine Einbindung von einer generativen KI in Software Systeme vor, um die Erklärbarkeit für den Nutzer zu steigern. In diesem Fall wird ChatGPT als generative KI genutzt. Der beim Nutzer aufkommende Erklärungsbedarf kann auf diese Weise im Software System selbst geklärt werden. Entsprechend sollten die Erklärungen prägnant ausfallen, da die Einbindung nur einen Teil der Oberfläche des Systems einnehmen sollte.

Der Nutzer gibt an einer designierten Stelle in der Softwareanwendung seine Frage ein, welche dann von der generativen KI beantwortet wird. Erklärungen können nach Bedarf passend zum aktuellen Kontext und Ziel angefordert werden. Auf vorherige Fragen und Antworten kann Bezug genommen und Nachfragen können gestellt werden.

Die Entwickler müssen vorher keine Erklärbarkeitsanforderungen erheben oder diese umsetzen, da die generative KI die Aufgabe der Bereitstellung von Erklärungen übernimmt. ChatGPT würde in diesem Szenario ohne vorherige Anpassungen in das Software System eingebunden.

Der Nutzungshintergrund bzw. Kontext kann sehr divers sein und somit auch die verwendete Software, für welche Erklärungen generiert werden sollen. Die Anwendung kann beispielsweise zum Zwecke der Arbeit, des Studiums oder privat in der Freizeit genutzt werden.

Die Stakeholder im Szenario sind die Anwender der Software, die diesbezüglich aber vielschichtig sein können. Der Nutzer kann sehr unerfahren bzw. neu in der Software sein oder bereits erste Erfahrungen gesammelt haben. Ebenfalls sind auch deutlich fortgeschrittene sowie Experten-Nutzer denkbar. Das technische Vorwissen kann im Generellen wie auch speziell bezüglich generativer KI schwanken. Es ist daher wichtig, dass sowohl erfahrene als auch unerfahrene Nutzer von den Antworten profitieren können.

Auf die in 2.1 aufgestellte Definition der Erklärbarkeit angewendet, bedeutet das also folgendes:

> Im Anwendungsszenario ist das System S ein beliebiges Software System. Der zu erklärende Aspekt X wird vom Nutzer, welcher den Adressaten A darstellt, im Kontext C bzw. Rahmen dieser Software geäußert. Der Aspekt X kann sich dabei auf verschiedene Arten von Erklärungsbedarf beziehen. In dieser Arbeit wurden Fragen zur *Interaktion*, dem *Systemverhalten*, der *Security* und dem *Domainwissen* näher betrachtet. Die Rolle des Erklärenden nimmt ChatGPT ein, welcher die Information I als Erklärung zu dem auf X basierenden Anfrageprompt gibt. Diese Erklärung soll dem Nutzer ermöglichen, die geäußerte Unklarheit zu verstehen.

3.1.2 Forschungsfragen

Die generierten Erklärungen sollen dem Nutzer das entsprechende Software System näherbringen, dessen Funktionen und Inhalte verständlich sowie schnell und einfach zugänglich machen.

Insgesamt ergeben sich vier Forschungsfragen.

RQ1:
Inwiefern kann eine generative KI wie ChatGPT prägnante Erklärungen zu verschiedenem Erklärungsbedarf und unterschiedlichen Software Systemen erzeugen?

RQ1 stellt das übergeordnete Ziel der Forschung dieser Arbeit dar. Ob eine Einbindung von generativer KI in ein Software System generell sinnvoll ist bzw. das manuelle Erstellen von Erklärungen ersetzen kann, hängt davon ab, ob der Erklärungsbedarf ausreichend gedeckt bzw. von der Qualität zufriedenstellend und prägnant ist.

RQ2:
Inwiefern beeinflussen die Erklärungen die Nutzung bzw. Bedienung und das Verständnis des Nutzers gegenüber der Software?

Die Bereitstellung und Inhalte der Erklärungen haben Auswirkung auf den Nutzer und sein Verhalten. Darunter fällt neben der Zufriedenheit des Nutzers, ob die Informationen korrekt oder richtungsweisend sind und dem Nutzer seine Arbeit erleichtern und effizienter gestalten können.

RQ3:
Welche Aspekte der Erklärbarkeit sind besonders wichtig für den Nutzer?

RQ3 untersucht, inwiefern die aufgestellten Metriken zur Bewertung der Erklärbarkeit für den Nutzer relevant sind. Beispielsweise könnte die Korrektheit der Antworten ein wichtiges Kriterium sein, auf die der Anwender besonders viel Wert legt. Womöglich reicht es dem Nutzer aber auch, eine ‚inhaltlich' falsche Antwort zu bekommen, welche für ihn einen Ansatz liefert, mit welchem er weiterarbeiten kann.

RQ4:
Welche weiteren Faktoren haben Einfluss auf die Güte der Erklärungen?

RQ4 klärt, inwieweit die Zufriedenheit des Nutzers außer von eigenen Faktoren bzw. deren Gewichtung auch von ‚externen' abhängig ist: Kategorisierung der Antworten, Art des Erklärungsbedarfs, Verbreitungsgrad der Software, welcher Einfluss auf die Wissensbasis für ChatGPT hat. Die erfassten Ergebnisse zu den erhobenen Daten werden auf signifikante Unterschiede untersucht.

Hypothesen der Arbeit

1. ChatGPT ist in der Lage, gute Erklärungen für verschiedenen Erklärungsbedarf und unterschiedliche Art von Nutzern zu generieren (RQ1, RQ4).

2. Die Aussagen von ChatGPT können als automatisierter Ersatz zur manuellen Entwicklung von statischen Erklärungen verwendet werden (RQ1).

3. ChatGPTs Antworten sind überwiegend korrekt oder mindestens richtungsweisend (RQ1).

4. Dem Nutzer wird durch die Antworten von ChatGPT die Bedienung der Software erleichtert (RQ2).

5. Der Arbeitsfluss wird nicht unterbrochen, sondern unterstützt (RQ2).

6. ChatGPT stellt Spekulationen an, die als solches auf Anhieb nicht erkennbar sind (RQ1, RQ2, RQ4).

7. Bei weniger bekannten oder verbreiteten Programmen sowie bei Fragen zu Programm-/Betriebsinterna wird ChatGPT Probleme haben (RQ1, RQ4).

8. Für den Nutzer ist die Korrektheit der Antworten wichtig, von ausschlaggebender Rolle ist jedoch die verständliche, richtungsweisende Qualität (RQ3).

9. Insgesamt wird die Software-Einbindung als hilfreich angesehen (RQ1, RQ2).

3.1.3 Vorgehensweise

Zur Klärung der Forschungsfragen und Prüfung der Hypothesen wurde ein Arbeitsablauf mit vier Hauptaspekten entworfen (Abbildung 3.1).

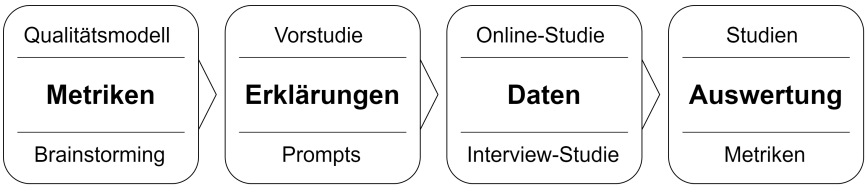


Abbildung 3.1: Vorgehen der Arbeit

Zunächst wurden anhand des Qualitätsmodells von Deters et al. [12] wichtige Aspekte und Kriterien zur Bewertung der Erklärbarkeit identifiziert. Darauf aufbauend wurde die Literatur auf weitere Ansätze und ergänzende Metriken untersucht. Außerdem wurden eigene Ideen gesammelt; es wurde insbesondere versucht, möglichst objektive bzw. messbare Metriken zu ergänzen. Schließlich wurden die Metriken zusammengetragen und auf ihren Zweck geprüft. Im nächsten Schritt wurden auf einer Vorstudie [16] aufbauend Prompts entwickelt, welche unterschiedlichen Erklärungsbedarf enthielten. Die Anfragen wurden an ChatGPT gestellt und Erklärungen generiert. Anschließend erfolgte eine quantitative Studie in Form einer eigens implementierten Online-Umfrage. In dieser wurden die subjektiven Metriken abgefragt. Es folgte zur Erweiterung und Validierung der ersten Ergebnisse eine qualitative Interview-Studie. Im Anschluss wurden die Daten der Studie wie auch die verbliebenen Metriken ausgewertet. Anhand der gesammelten Ergebnisse wurden die generierte Erklärbarkeitsqualität von ChatGPT bewertet und die Forschungsfragen evaluiert.

3.2 Metriken

Durch die automatisierte Erstellung von Erklärungen ist die Qualität der Erklärungen von der generierenden KI abhängig, in diesem Fall ChatGPT. Anhand des Qualitätsmodells von Deters et al. [12] werden verschiedene, im Anwendungsszenario anstrebenswerte Aspekte der Erklärbarkeit betrachtet und die Erklärungen daraufhin untersucht.

3.2.1 Aspekte der Erklärbarkeit

Im Qualitätsmodell [12] werden zehn Aspekte der Erklärbarkeit aufgestellt:

A1: Verständlichkeit, A2: Transparenz, A3: Effektivität,
A4: Effizienz, A5: Zufriedenheit, A6: Korrektheit,
A7: Angemessenheit, A8: Vertrauenswürdigkeit,
A9: Überzeugungskraft, A10: Überprüfbarkeit

Im Folgenden wird die Relevanz der einzelnen Aspekte für das beschriebene Anwendungsszenario (3.1.1) eingeordnet. Dabei ergibt sich aus der Zieldefinition als zentraler Aspekt die Zufriedenheit (A5) des Nutzers durch die Bereitstellung der generierten Erklärungen. Ausgehend davon wird die Wichtigkeit der anderen Aspekte eingestuft.

Mit der Zufriedenheit eng verbunden ist die Verständlichkeit (A1). Damit eine Erklärung zufriedenstellend sein kann, muss sie zunächst verstanden werden können. Außerdem sollte die Frage des Nutzers sinnvoll beantwortet werden und die Erklärung ihm weiterhelfen. Diese Nützlichkeit wird unter dem Punkt der Effektivität (A3) näher untersucht. Als weiterer wesentlicher Aspekt wird die Korrektheit (A6) der generierten Antwort gesehen. Zum einen kann es frustrierend sein, wenn der Nutzer Anweisungen befolgt, welche sich als nicht korrekt herausstellen. Zum anderen sollten keine falschen Informationen über das System verbreitet werden, auch wenn der Nutzer die Korrektheit nicht selbst nachvollziehen kann. Ziel ist es, manuell erstellte Antworten unabhängig von der Zufriedenstellung des Nutzers ersetzen zu können, ohne an Qualität zu verlieren (RQ1).

Abhängig vom Erklärungstyp können außerdem Angemessenheit (A7) und Effizienz (A4) Auswirkungen auf die Zufriedenheit des Nutzers haben. Wie-Erklärungen könnten ein schnelleres Arbeiten bewirken. Eine zu ausführliche oder auch generelle Begriffserklärung dagegen trägt unweigerlich weniger zur Effizienzsteigerung bei. Die Erklärungen sollten generell angemessen an die Bedürfnisse des Nutzers und angepasst an das System ausfallen. Dies sollte insbesondere bei Was- und Warum-Fragen der Fall sein. Bei diesen wird nach einer Erläuterung gefragt, deren Ausführlichkeit entsprechend adäquat sein sollte.

Im Anwendungsszenario wird kein spezieller Fokus auf Transparenz (A2) und Vertrauen (A8) des Nutzers gegenüber der erklärten Software gelegt. Dennoch stellen sie Nebenaspekte dar, die in die Qualität der Erklärbarkeit mit einbezogen werden können. Stellt der Nutzer eine Frage, welche das System näher erklären soll, sollte das System transparent gemacht werden. Gleichzeitig kann eine zufriedenstellende Antwort das Vertrauen in die erklärte Software steigern. Näher untersucht wird das Vertrauen in ChatGPT.

Nicht weiter betrachtet werden die Aspekte Überzeugungskraft (A9) und Überprüfbarkeit (A10). Es ist nicht der Fokus des Szenarios, den Nutzer von etwas zu überzeugen oder ihn zu veranlassen, etwas Bestimmtes zu tun. Es geht vielmehr um die Bereitstellung der Erklärungen selbst. Ebenso liegt der Fokus nicht darauf, das System für den Nutzer überprüfbarer und damit zusammenhängend korrigierbarer zu gestalten.

Damit ergeben sich die in Abbildung 3.2 dargestellten zentralen Aspekte zur Bewertung der Erklärbarkeit der Software Systeme durch ChatGPT sowie die Nebenaspekte Transparenz und Vertrauen. Die zentralen Aspekte sind mit einem grauen Balken hinterlegt. Die darunterstehenden Aspekte stehen in Bezug zu bestimmten Erklärungstypen und werden dahingehend spzifizierter betrach.

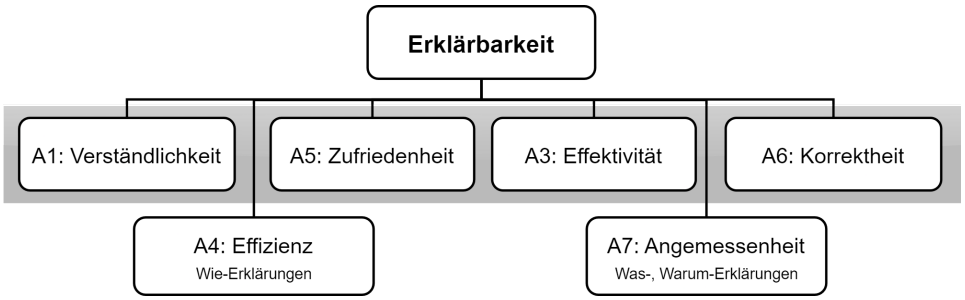


Abbildung 3.2: Wichtige identifizierte Aspekte zur Bewertung der Erklärbarkeit [12] im Anwendungsszenario

3.2.2 Kriterien und Metriken der Aspekte

Im Folgenden werden die einzelnen Aspekte, welche als besonders wichtig erachtet wurden, näher betrachtet. Die Kriterien wurden, ebenso wie die Aspekte, dem Qualitätsmodell [12] samt Nummerierung entnommen und für die hier vorliegenden Zwecke angepasst. Darauf basierend wurden entsprechende Metriken zusammengestellt. Die Metriken sind nicht direkt aus dem Qualitätsmodell übernommen, sondern entsprechend adaptiert und erweitert worden.

Zufriedenheit stellt den Hauptaspekt der Untersuchung dar und zielt insbesondere auf RQ1 ab: Das Potenzial, Erklärungen zu generieren, steht im starken Zusammenhang damit, wie zufrieden der Nutzer mit diesen ist. Die Zufriedenheit ist abhängig von mehreren Faktoren. Neben der Erfüllung der anderen Qualitätsaspekte kann die Realisierung der Zufriedenheit in die in Abbildung 3.3 abgebildeten Kriterien aufgeteilt werden.

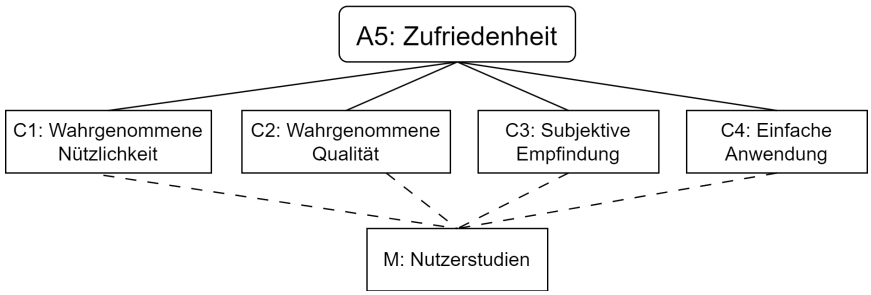


Abbildung 3.3: ausgewählte Kriterien und Metriken Zufriedenheit

Da die Zufriedenheit des Nutzers subjektiv empfunden wird, müssen die wahrgenommene Nützlichkeit, Qualität, Empfindung und Einfachheit der Ver- bzw. Anwendung der Erklärung durch Nutzerstudien erfasst werden.

Das Kriterium C5.5 Einfachheit wird hier nicht extra gelistet, sondern nur im Rahmen des Aspekts der Verständlichkeit betrachtet, da hierunter genau diese Einfachheit der Sprache verstanden wird.

Verständlichkeit Die Erfassung des Aspektes der Verständlichkeit gestaltet sich komplex. In der Literatur gibt es verschiedene Ansätze (vgl. Abschnitt 2.5). Als häufig verwendeten Lesbarkeitsindex wird die auf die deutsche Sprache angepasste Variante des Flesch Reading Ease Scores angewendet. Weiter werden eigene Metriken aufgestellt (vgl. Abbildung 3.4).

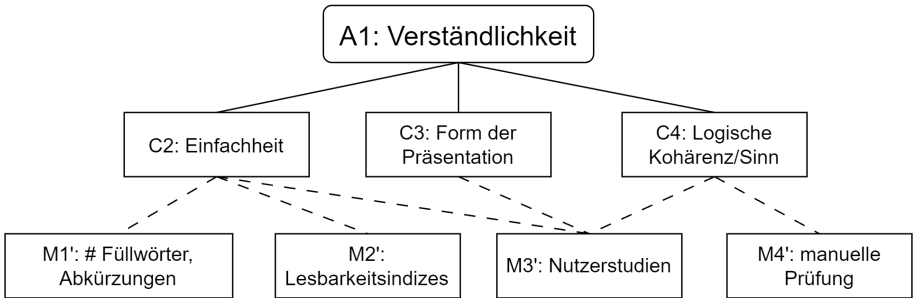


Abbildung 3.4: Ausgewählte Kriterien und Metriken Verständlichkeit

Das Kriterium C1.1 Kognitive Beanspruchung wurde nicht untersucht. Die Erfassung über beispielsweise den im Qualitätsmodell vorgeschlagenen NASA-TLX cognitive load score würde einen erheblichen Teil der Studie beanspruchen. Die Anzahl an Fragen im Studienrahmen dieser Arbeit sind begrenzt, daher beschränken sie sich auf die wichtigsten Metriken über alle Aspekte hinweg.

Für die Einfachheit der Erklärungen werden verschiedene Ansätze verfolgt. Zum einen wird der Lesbarkeitsindex nach Flesch [18] ermittelt. Des Weiteren werden die je höchstgewichteten Variablen mit negativem sowie

positiven Einfluss auf die Verständlichkeit des KWM [3] betrachtet. Die Satzlänge (negativer Einfluss) und Verbintensität (positiver Einfluss) wurden als ausschlaggebende Punkte des KWM angesehen. Andere Variablen würden ferner eine sehr genaue Betrachtung und manuelle Analyse zur Bestimmung ihrer Anteile benötigen.

Bei Füllwörtern ist ein geringes Vorkommen positiv zu bewerten, da sie in der Regel den Text aufblähen, ohne inhaltlichen Beitrag zu leisten. Abkürzungen könnten das Verständnis des Textes erschweren, daher sind wenige Abkürzungen in den Erklärungen ebenfalls positiv bewertet. Die Form der Erklärung ist durch ChatGPT eingeschränkt. In der verwendeten Version werden keine Bilder, Abbildungen oder Ähnliches erstellt, welche zum Verständnis positiv beitragen würden [11]. Allerdings können Tabellen, nummerierte Listen und mit ‚<' versehene Abfolgen generiert werden, deren Art des Einflusses zu klären ist. Die Form wird als Nebenkriterium wie die anderen beiden Kriterien in den Nutzerstudien mituntersucht.

Unter dem Kriterium der logischen Kohärenz ist eine klare Struktur der Erklärungen positiv zu bewerten. Außerdem sollten die Aussagen keine Widersprüche enthalten. Es wird abgefragt, ob die generierten Erklärungen für den Nutzer Sinn ergeben. Außerdem wird eine manuelle Prüfung durch den Autor durchgeführt.

Effektivität Die Erklärungen von ChatGPT sollen hilfreich sein und den Nutzer darin unterstützen, das erklärte Software System besser zu verwenden.

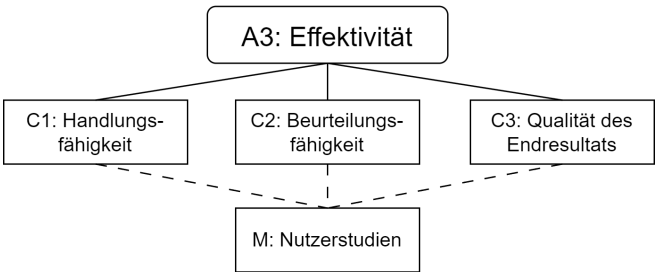


Abbildung 3.5: Ausgewählte Kriterien und Metriken Effektivität

Die in Abbildung 3.5 abgebildeten Kriterien werden mittels der Nutzerstudien untersucht, primär in der Interview-Studie. Die Handlungsfähigkeit bezieht sich darauf, eine Aktion durchzuführen bzw. dazu in der Lage zu sein. Die Teilnehmer können die Erklärungen direkt ausprobieren und anschließend die Qualität des Resultats der Anwendung der Erklärung einstufen. Währenddessen kann beobachtet werden, inwieweit der Nutzer in der Lage ist, die Situation zu beurteilen. In der Online-Studie wird allgemein nach der wahrgenommenen Effektivität gefragt, also, ob der Nutzer die Erklärung hilfreich findet.

Korrektheit Die Antworten von ChatGPT sollten korrekt sein und keine falschen Informationen geben. Es werden die in Abbildung 3.6 dargestellten Kriterien näher untersucht.

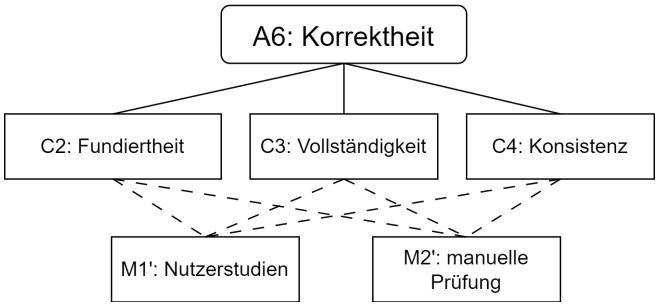


Abbildung 3.6: Ausgewählte Kriterien und Metriken Korrektheit

Das erste Kriterium C1 Aufrichtigkeit bezieht sich darauf, den Nutzer nicht absichtlich in die Irre zu führen. Es ist davon auszugehen, dass bei generierten Erklärungen keine Absicht in diesem Sinne vorliegt. Daher wird die Aufrichtigkeit nicht näher betrachtet. Die Antworten sollten inhaltlich korrekt, zuverlässig und stichhaltig sein. Ein Unterpunkt der Korrektheit ist die Vollständigkeit. Die Erklärungen sollten konsistent und in sich übereinstimmend sein. Ähnliche Prompts sollten gleiche oder ähnliche Ergebnisse erzielen.

Alle Kriterien werden näher in der Interview-Studie beleuchtet. In der Online-Studie gibt es bezüglich der Korrektheit keine direkte Abfrage. Der Nutzer hat die Möglichkeit, Auffälligkeiten in den Kommentaren zu äußern. Neben der Bewertung der Konsistenz innerhalb einer Erklärung durch den Nutzer wird die Konsistenz zu einer alternativ generierte Erklärung bewertet. In der manuellen Überprüfung erfolgt eine Einteilung der Aussagen in falsche, ‚halbrichtige' und nicht belegbare Aussagen, bei denen die Korrektheit angezweifelt werden kann. ‚Halbrichtige' Aussagen weisen in die richtige Richtung, enthalten aber beispielsweise abweichende Begrifflichkeiten im Vergleich zu den in der Software vorhandenen.

Effizienz Die Erklärungen sollen den Nutzer befähigen, seine Ziele schneller zu erreichen bzw. die Software und ihre Funktionen schneller benutzen zu können. Aus diesem Grund wird in der Online-Studie das Kriterium ausschließlich bei den Wie-Fragen abgefragt, welche in der Regel eine Anleitung anfordern.

In Abbildung 3.7 sind die wichtigsten identifizierten Kriterien der Effizienz dargestellt. Die schnellere Nutzung, die Zeit zum Zugang zu den Informationen werden über die Studien ermittelt. C3, die Abwesenheit von irrelevanten Informationen, wird nicht betrachtet. Der zeitlich gesehene Zugang zu Informationen wurde für das Anwendungsszenario dieser Arbeit

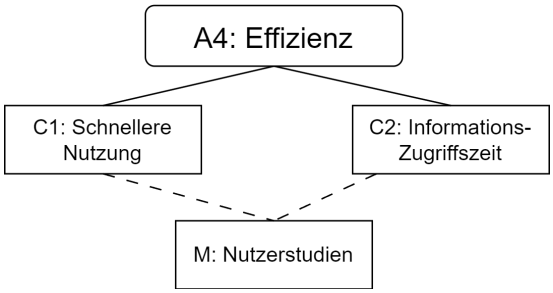


Abbildung 3.7: Ausgewählte Kriterien und Metriken Effizienz

interpretiert: ChatGPT beantwortet die Frage direkt und liefert benötigte Informationen, anstatt auf eine andere Quelle verweist. Die benötigte Zeit für den mentalen Zugang der Erklärung konnten in der Online-Studie nicht direkt gemessen werden. Hier wäre es nicht klar und überprüfbar, wann der Teilnehmer beginnt, die Frage sowie die erklärende Antwort zu lesen, und wie lange er dann braucht, sie zu verarbeiten. In der Interview-Studie wurden die benötigten Zeiten für den jeweiligen Abschnitt grob festgehalten. Die Generierungsgeschwindigkeit von ChatGPT wurde nicht gemessen; in der Online-Studie wurden die Zeiten vorgeneriert zur Verfügung gestellt. In der Interview-Studie wurden bezüglich der einzelnen Kriterien keine direkten Fragen gestellt, Äußerungen des Teilnehmers aber mit aufgenommen.

Angemessenheit Die Angemessenheit stellt in der Online-Studie das Gegenstück zur Effizienz dar und wird bei den Warum- und Was-Fragen abgefragt. Diese legen den Fokus auf ein inhaltliches Verständnis, bei der die Angemessenheit eine größere Rolle spielt.

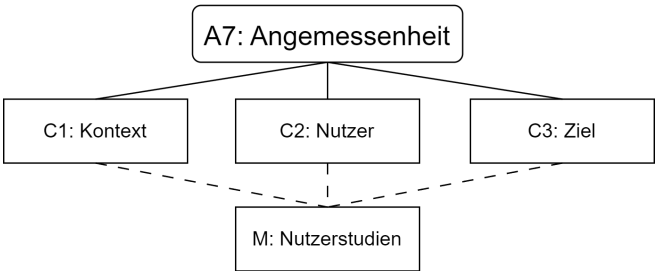


Abbildung 3.8: Ausgewählte Kriterien und Metriken Angemessenheit

Abbildung 3.8 zeigt die einzelnen Kriterien: Erklärungen können in Bezug auf den Kontext, den Nutzer und das angestrebte Ziel angemessen sein. Ob die Frage des Nutzers ausreichend beantwortet wurde, wird in den Nutzerstudien erfasst. Außerdem kann der Ton der Sprache mittels einer Sentimentanalyse ermittelt werden. Ein neutraler bis positiver Ton mit objektiver Haltung sind positiv bewertet.

Transparenz Im Anwendungsszenario selbst steht das Transparentmachen des Systems nicht im Vordergrund; der Fokus auf diesen Aspekt ist abhängig von der Frage des Nutzers. Durch die Erklärungen sollen die Software Systeme dem Nutzer nähergebracht werden. Das mentale Modell des Nutzers und die Vorhersagbarkeit seiner Aktionen werden durch die Erklärungen womöglich verbessert, auch wenn der Nutzer mit seiner Frage nicht darauf abzielte. Die im Qualitätsmodell vorkommenden Kriterien werden hier daher eher oberflächlich untersucht. ChatGPT sollte System Komponenten, Funktionen und Elemente so darlegen können, wie es vom Nutzer angefragt wird (Angemessenheit). Die Auswirkungen auf die wahrgenommene Transparenz des Systems werden in den Nutzerstudien abgefragt.

Vertrauen Ähnlich ist es bei dem Aspekt des Vertrauens des Nutzers in das jeweilige Software System. Für das Anwendungsszenario wird es als weniger wichtig angesehen, dass das Vertrauen des Nutzers in die erklärte Software gesteigert wird. Vielmehr wird hier der Fokus darauf gelegt, inwieweit der Nutzer Vertrauen zu den generierten Antworten bzw. ChatGPT selbst hat. Hier wurde primär das Vertrauen der Nutzer in Hinblick auf die durch ChatGPT generierten Antworten bzw. in ChatGPT selbst untersucht. Entsprechend werden die Erklärungen selbst nicht nach dem im Qualitätsmodell [12] definierten Aspekt Vertrauen untersucht.

3.2.3 Ideen anderer Metriken

Um für die einzelnen Aspekte und Kriterien passende Metriken aufzustellen, wurden zunächst eigene Ideen gesammelt. Außerdem wurde ChatGPT abschließend zur Ergänzung nach Metriken gefragt.

Verworfene Ansätze

Für die Bewertung der Einfachheit wurde überlegt, die Anzahl längerer Wörter zu zählen. Hier war jedoch unklar, ab welcher Länge das Wort einen Satz schwerer zu lesen macht. Außerdem war eine Idee, die Anzahl Stoppwörter zu zählen. Stoppwörter (z.B. *und*, *oder*, *der*, *die*, *das*) sind häufig vorkommende Wörter, welche wenig zur eigentlichen Bedeutung des Textes beitragen, und sollten daher leicht verständlich sein. Sie haben Einfluss auf die Länge des Satzes, welche bei der Berechnung der Lesbarkeitsindizes eine wichtige Rolle spielt. Gleichzeitig sind Stoppwörter meist kurz und somit von geringer Silbenzahl, welche ebenfalls großen Einfluss auf den Flesch Reading Ease Score nimmt. Daher wurde entschieden, Stoppwörter nicht weiter gesondert zu betrachten.

Es wäre denkbar, weitere Formeln zur Berechnung des Lesbarkeitsindexes zu verwenden. Hier wurde sich auf den Flesch Reading Ease Score und das KWM zur Abdeckung beschränkt. Bei einem Vergleich von Briest [3] erzielten andere Varianten keine signifikant besseren Ergebnisse. Immel [23] zweifelte zudem die Aussagekraft solcher Formeln zur Bewertung der Verständlichkeit an.

Einfluss auf die Einfachheit könnten auch die Verwendung technischer Begriffe, Metapher und irrelevante Informationen bzw. streichbarer Wörter haben. Aufgrund der hohen Anzahl generierter Antworten stünde eine Überprüfung dahingehend nicht im Verhältnis zur erhaltenen Aussage.

Die generelle Nutzbarkeit könnte durch Usability Tests [42] und Metriken näher untersucht werden. Dies ist jedoch ein spezieller Aspekt, welcher nicht im Hauptfokus der Forschungsfragen stand.

Vorschläge von ChatGPT

Bei der zusätzlichen, im Anschluss durchgeführten Befragung von ChatGPT stellte dieser acht Kategorien auf, mit je zwei Unterkategorien und möglichen Metriken. Gefragt wurde nach Metriken zur Bewertung von kurzen Erklärungen.

Die Erste lautete *‚Lesbarkeit und Struktur‘* mit den Unterpunkten *Lesbarkeitsindizes* und *Strukturanalyse*. Der erste Punkt war bereits durch Metriken abgedeckt. Für den zweiten Punkt wurden Textzusammenfassungsalgortihmen vorgeschlagen. Da die Erklärungen von ChatGPT im angestrebten Anwendungsszenario recht kurz ausfallen sollen, scheint dies in diesem Kontext nicht sinnvoll. Ähnliches galt für die Punkte der Kategorie *‚Kürze und Prägnanz‘*. Auch *‚Überzeugungskraft und Argumentation‘* sowie die *‚inhaltliche Relevanz‘* sind für die angestrebte Kürze der Erklärungen nicht sinnvoll mittels Berechnungen zu analysieren. Eine darauf basierende eigene Idee wäre, die Ähnlichkeit zwischen generierten Erklärungen näher bezüglich ihrer Konsistenz zu untersuchen. Die Konsistenz ist bereits indirekt in der Nutzerstudie von den Teilnehmern zu beurteilen. In dieser Arbeit wird sich auf zwei ChatGPT Antworten beschränkt und die Konsistenz in den Bewertungen betrachtet. Weitere Kategorien sind *‚Wortwahl und Stilistik‘*, *‚Publikumsorientierung‘* und *‚Feedback und Reaktionen‘*, welche die im Folgenden erläuterten Punkte enthielten. Eine *Sentimentanalyse* wird durchgeführt. Die *Stilometrie* wird nicht näher betrachtet, da auch hier die Erklärungen zu kurz für eine Analyse von Wortfrequenzen und -mustern sind. Satzlängen werden durch die Lesbarkeitsindizes abgedeckt. *Social-Media-* und *Zielgruppenanalyse* werden nicht durchgeführt. Es werden Rückmeldungen und Reaktionen aus den Umfragen berücksichtigt. *‚Grammatik und Rechtschreibung‘* wurden nicht untersucht.

3.3 Fragen für die Online-Studie

Die in Unterabschnitt 3.2.1 identifizierten Aspekte sind Verständlichkeit, Zufriedenheit, Effektivität und Korrektheit. Hinzu kommen abhängig vom Erklärungstyp der Antworten die Effizienz und Angemessenheit, welche als ‚kategoriespezifische' Frage zusammengefasst werden können. Abgesehen von der Korrektheit sollten alle Aspekte in der Nutzerstudie direkt abgefragt werden. Der Nutzer kann und soll die Korrektheit der Antwort auf die Frage nicht überprüfen, da dies zum einen den zeitlichen Rahmen der Studie ausdehnen würde und zum anderen außerhalb des Rahmens seiner Möglichkeiten liegt.

Die geplante Online-Studie soll in erster Linie viele Beispiele von Nutzerfragen und Antworten von ChatGPT enthalten. Aus Machbarkeitsgründen musste deshalb die Anzahl Bewertungsfragen an den Nutzer pro Frage-Antwort-Paar begrenzt werden. Zu jedem Aspekt wird sich daher auf eine Bewertungsfrage beschränkt. Zunächst wurden mehrere mögliche Fragen bzw. zu bewertende Aussagen pro Aspekt aufgestellt und diese anschließend priorisiert. Die ausgewählte Frage sollte den Aspekt und die unterteilten Kategorien erfassen und in dem Kontext einer Online-Umfrage beantwortbar sein. Vor allem sollten die subjektiven Wahrnehmungen der Nutzer eingefangen werden. Alle Fragen sind im Fragebogen als Aussagen formuliert, zu denen der Nutzer seine Zustimmung angeben soll:

- Zufriedenheit: *„Insgesamt bin ich zufrieden mit der Erklärung"*
  Die Zufriedenheit kann direkt abgefragt werden: Wie zufrieden ist der Nutzer mit der Erklärung. Die Kriterien der Zufriedenheit werden zusätzlich durch die Bewertung der anderen Aspekte mit abgedeckt.

- Verständlichkeit: *„Die Erklärung ergibt Sinn für mich"*
  Bei der Verständlichkeit könnte nach der Struktur bzw. Form der Präsentation oder auch der Länge der Antworten gefragt werden. Die Länge der Erklärung kann bei der Angemessenheit eine Rolle spielen, ebenso wie die Struktur. Daher sollte der Nutzer allgemein beurteilen, ob die Erklärung für ihn Sinn ergibt.

- Effektivität: *„Ich finde die Erklärung hilfreich"*
  Die Effektivität soll festhalten, ob die Erklärung für den Nutzer hilfreich ist. Genau das ist der Gegenstand der Beurteilung durch den Nutzer.

- Effizienz: *„Die Erklärung ermöglicht mir schnelleres und/oder einfacheres Arbeiten"*
  Ebenso eindeutig ist die Effizienz; der Nutzer soll einschätzen, ob ihm durch die (Wie-)Erklärung ein schnelleres Arbeiten ermöglicht wird.

- Angemessenheit: *„Die Frage ist für mich ausreichend/zufriedenstellend beantwortet“*
  Als Alternative zur Effizienz soll beurteilt werden, ob die Angemessenheit ausreichend bzw. zufriedenstellend ist, ohne diese näher zu spezifizieren. Der Nutzer bestimmt, welche Kriterien für ihn dabei eine Rolle spielen.

Als Nebenaspekte wurden die Transparenz(-steigerung) der erklärten Software und das Vertrauen in die Antworten von ChatGPT identifiziert. Diese werden rückblickend vom Nutzer nach der Bewertung der einzelnen Frage-Antwort-Paare evaluiert. Der Nutzer kann dann beurteilen, ob die Erklärungen die einzelnen Software Systeme für ihn transparenter bzw. verständlicher gemacht haben. Die Transparenz wird nicht bei jedem Frage-Antwort-Paar abgefragt, da sie von der Fragestellung und Intention des Fragestellenden abhängt. Er hat eigene Erfahrungen mit Antworten von ChatGPT sammeln können und kann daraufhin bewerten, wie vertrauenswürdig ihm diese erscheinen. Vergleichend wird der Wert zu Beginn der Umfrage abgefragt, um eine mögliche Änderung im Vertrauen zu erfassen.

Kapitel 4

Entwicklung der Umfrageinhalte

Ausgangspunkt für die Entwicklung möglicher Anfragen an ChatGPT war eine Studie zur Erfassung von Erklärungsbedarf von Droste et al. [16]. Die dort gesammelten Unklarheiten wurden mit dem Ziel weiterverarbeitet, Fragen zu extrahieren und Prompts abzuleiten. Darauf basierend wurden Antworten von ChatGPT-4 generiert und die Fragen und Antworten auf die Verwendung in der Nutzerstudie vorbereitet.

4.1 Promptentwicklung

In der Studie von Droste et al. [16] wurden Fragen und Unklarheiten des Nutzers zu zuletzt genutzter Software gesammelt. Erfasst wurden hier unter anderem die Arten zum Erklärungsbedarf zu Domainwissen, Systemverhalten, Security und Interaktion. In diesem Abschnitt werden diese Unklarheiten auch als Antworten betitelt, da sie nicht immer eine Frage darstellten und als Antwort auf die von Droste et al. durchgeführte Studie angesehen wurden. Die geäußerten Unklarheiten waren unterteilt in unterschiedlichen Arten von Erklärungsbedarf, wie zum Beispiel zum Systemverhalten und der Sicherheit des Systems. Insgesamt wurden 315 Antworten zu über 75 Software Systemen erfasst.

Die Daten wurden zur Promptentwicklung in drei Iterationen weiterverarbeitet (Abbildung 4.1). In der ersten Iteration erfolgt eine grobe Vorsortierung der Daten. Hier werden nicht verwendbare Äußerungen aussortiert und jene behalten, welche eine direkt als Prompt verwendbare Frage darstellten. In der zweiten Iteration wurden die noch nicht gelabelten Daten weiter untersucht. Nicht eindeutig einsortierbare Aussagen wurden gekennzeichnet. Außerdem wurde ein Label vergeben, welches diese Einordnung näher ausführt. In der dritten und letzten Iteration wurden schließlich verbliebene Aussagen der Nutzer zu Fragen umformuliert. Konnte keine konkrete Frage

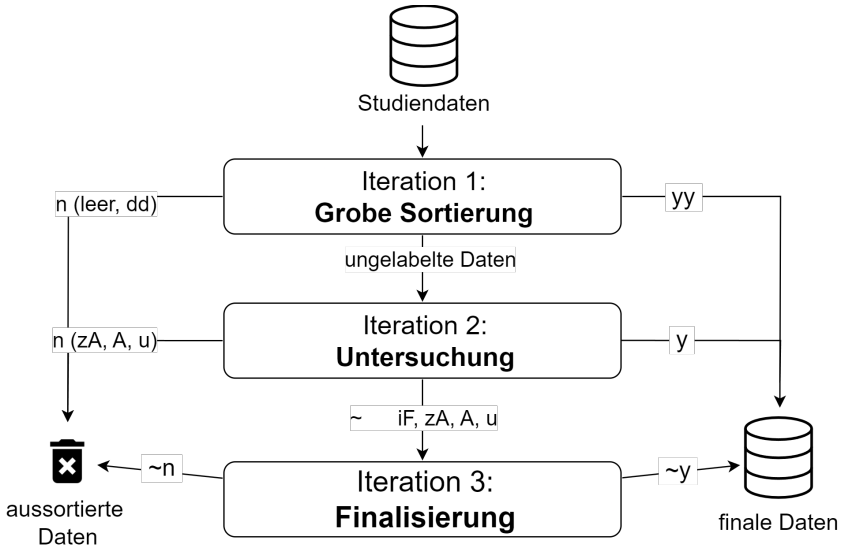


Abbildung 4.1: Iterationen zur Entwicklung der finalen Daten

aus den Unklarheiten gezogen werden, wurden diese verworfen. Schließlich wurden an den finalen Daten letzte Anpassungen für die Verwendung als Prompt vorgenommen. Eine Übersicht der verwendeten Label findet sich in Tabelle 4.1. Auf diese wird in den Beschreibungen der einzelnen Iterationen näher eingegangen.

Iteration 1: Grobe Sortierung Im ersten Schritt wurden die Daten zunächst zusammengeführt, in ein einheitliches Format gebracht und grob kategorisiert. Die zur Bearbeitung verfügbaren gestellten Arten von Erklärungsbedarf waren *Domainwissen*, *Interaktion*, *Security* und *Systemverhalten*. Neben einer vorgegebenen Antwort-ID, welche mehrfach vorkommen konnte, der Angabe der Software und der Art des Erklärungsbedarfes war der Erklärungsbedarf selbst festgehalten.

Da die Daten nicht vollständig zur Verfügung gestellt waren, gab es zu manchen IDs keine Einträge. Insgesamt 34 leere Datensätze wurden aussortiert (n, leer). Teilweise gab es verschiedene IDs mit gleichen oder selben Unklarheiten der Nutzer. Zum Beispiel wurde zu der Software *Word* unter dem Erklärungsbedarf *Interaktion* mehrmals danach gefragt, wie es möglich ist, Seitenzahlen erst ab Seite 3 anzeigen zu lassen. Davon wurde je ein Eintrag behalten und insgesamt 16 unterschiedliche Doppelungen aussortiert (n, dd). Wenn zu einer Antworten-ID mehrere Arten von Erklärungsbedarf bzw. Fragen auftauchten, wurden diese getrennt voneinander behandelt. Die übernehmbaren Fragen *„Wie benutzt man den Arbeitsplatz?“* und *„Wie benutzt man die Unterrichtsplanung?“* zu der Software *Studip* waren unter der selben ID gelistet.

Auf diese Weise entstandenen knapp 300 Datensätze, welche auf ihre direkte Eignung als Anfrage für ChatGPT untersucht wurden (yy). Behalten wurden Einträge, welche eine konkrete Frage darstellten oder enthielten. Direkt übernehmbare Fragen waren beispielsweise *„Was ist OER Campus?“* bei der Software *Studip* oder *„Wie sicher ist Telegramm?“*. Leicht umformuliert wurden Antworten wie *„Ob ich mein online Status verbergen kann.“* zu der Software *WhatsApp* oder *„Dateien teilen“* bei *GoodNotes*, sodass diese in Form einer Frage gestellt sind.

Label	**Beschreibung**
yy	Klares Behalten (teilweise direkt übernehmbar)
y	Behalten
~y	Nachträglich behalten (Konkretisierung einer Anfrage)
n	Nicht behalten
~n	Nachträglich aussortiert (keine konkrete Frage ableitbar)
iF	Indirekte Frage
zA	Zu bzw. sehr allgemein formuliert
A	Aussage, Meinung des Nutzers
u	Unklar
dd	Doppelt
leer	Leere Einträge

Tabelle 4.1: Beschreibung der Labels für das Behalten von Daten (oben) sowie der ausführenden Begründung für die Einordnung (unten)

Iteration 2: Untersuchung In der zweiten Iteration wurden die noch ungelabelten Daten als zu-behalten markiert (y), wenn sie offensichtlich zu einer Frage umformuliert werden konnten. Dies wurde in der ersten Iteration nur stellenweise getan; in der zweiten Iteration wurde nun der Fokus darauf gelegt. Daher sind Beispiele ähnlich zu denen in der ersten Iteration umformulierten Einträgen. *„Wo ich Einstellungen zu einem Server finde.“* bei der Nutzung von *Discord* oder *„Erstellen von Meetings,“* bei *Outlook* wurden umformuliert zu *„Wo finde ich bei Discord die Einstellungen zu einem Server?“* und *„Wie kann ich bei Outlook Meetings erstellen?“*.

Waren die Antworten deutlich zu allgemein gehalten (zA), unklar (u) oder eine Aussage oder Meinung des Nutzers (A) ohne (in-)direkte Frage wurden sie aussortiert (n) und der Grund festgehalten:

- Zu Allgemein (zA): Die *„Menüführung“* beim Herd und *„Nicht anklickbare features“* bei *League of Legends* wurden als zu allgemein eingestuft. Das entscheidende Kriterium ist, dass die Formulierungen unpräzise sind. Es konnte kein konkreter Prompt entwickelt werden, da ChatGPT zu umfassend antworten könnte.

- Unklar (u): Bei *„Manchmal durch Popups am kleinen Brave-Symbol oben rechts in der Adresszeile.“* oder *„Recherche zu fehlgeschlagenen updates“* (*Google Chrome*) war unklar, welche Frage genau der Nutzer hat. Eine Beurteilung der Korrektheit der generierten Antwort so wie das Geben einer speziellen Antwort wird als nicht sinnig bzw. möglich angesehen.

- Aussage (A): Als Aussagen des Nutzers wurden beispielsweise *„Vorstellung und Umsetzung weichen voneinander ab“* (*Shortcut*) und *„Manche Aussprachehilfen passen nicht zu den angezeigten Fremdwörtern oder sind verzerrt.“* (*Duolingo*) angesehen. Das Kriterium ist, dass hier die Formulierung einer Frage rein spekulativ und ggf. nicht zielführend ist.

Insgesamt waren viele Antworten zu ungenau, um daraus einen konkreten Prompt mit spezifizierten Kontextbezug und Antwortmöglichkeit herauszuziehen. Alle verbliebenen Daten wurden mit einer (~) markiert und in die dritte Iteration zur näheren Untersuchung weitergegeben. Außerdem wurde der Grund für die Einordnung angegeben. Dieser war jedoch nicht so stark ausgeprägt, wie bei den eben genannten Beispielen, sodass potenziell konkrete Prompts daraus entwickelt werden könnten.

Als weiteres Label wurde dazu (iF) für indirekte bzw. interpretierte Fragen eingeführt. Hier war entweder ein Teil der Nutzerantwort eine Frage oder es konnte eine Frage hineininterpretiert werden. In diesem Fall ist nicht sichergestellt, dass die eigens formulierte Frage dem intendierten Erklärungsbedarf des Nutzers entspricht. Allerdings basiert sie auf der Unklarheit des Nutzers und nimmt darauf Bezug. Ein Beispiel wäre die Aussage zu *Word „Formatierung -> wenn man z.B. ein Bild verschiebt, verschiebt sich teilweise alles“*, aus welcher folgende Frage abgeleitet werden könnte: *„Wie kann ich die aktuelle Formatierung in Word beibehalten, wenn ich ein Bild verschiebe? (in Microsoft Word)“*. Weitere Beispiele finden sich im nächsten Schritt.

Iteration 3: Finalisierung In der dritten und letzten Iteration wurden die mit (~) gelabelten Einträge unter folgenden Umständen als nachträglich zu-behalten markiert (~y):

- Nur mit (iF) gekennzeichneten Unklarheiten der Nutzer. Diese wurden zu konkreten Fragen umformuliert. Hierunter fielen beispielsweise die Änderung von *„Ich finde nicht auf Anhieb, welche User die Subscription beinhaltet“* (*Azure Portal*) in die Frage *„Wo finde ich beim Azure Portal, welche User die Subscription beinhaltet?“* oder von *„Die Flut an Smileys ist z.T verwirrend, wenn man einen ganz bestimmten finden will (z.B. Konfettikanone).“* (*WhatsApp*) zu *„Wie kann ich ein Emoji wie die Konfettikanone bei Whatsapp in der Flut an Smileys finden?“*.

- Aussagen (A), die als Fragen interpretiert werden konnten (iF). Beispiele hierzu sind *„Ja, auf Fakten ist keinerlei Verlass, teils völlige ‚Phantasieprodukt'."* (*ChatGPT*) zu *„Worauf basieren deine Informationen? Erzählst du Fakten oder ‚Phantasieprodukte'? In wie weit ist auf deine Fakten Verlass?"* und *„wenn eine Tabelle direkt am Anfang der Seite ist, ist es nicht so leicht möglich, darüber eine normale Zeile (z.B. für die Überschrift) einzufügen"* bei *Word* zu *„Gibt es einen Trick, wie ich bei Word, wenn am Anfang der Seite eine Tabelle ist, noch eine normale Zeile darüber einfügen kann?"*.

- Zu allgemeine Aussagen (zA), welche jedoch mit konkreten Beispielen von ChatGT beantwortet werden könnten: *„manchmal wird eine Benachrichtigung angezeigt, man kann aber nicht nachvollziehen, wo diese her kommt"* (*Instagram*) zu *„Bei Instagram werden Benachrichtigungen angezeigt, kannst du mir erklären, wo diese herkommt oder sie zu finden ist?"*.

- Unklare Aussagen (u), welche Interpretationsspielraum ließen: *„Bei der Nachrichtenfunktion Antwort gelöscht?"* (*Ebay Kleinanzeigen*) zu *„Kann ich bei der Nachrichtenfunktion von Ebay Kleinanzeigen Antworten (ausversehen) löschen? Kann ich das rückgängig machen?"*.

War keine Konkretisierung möglich (siehe auch Iteration 2) und konnte keine Frage erstellt werden, wurde der Datensatz aussortiert (~n).

Aus den behaltenen Daten wurden abschließend die verbliebenen Prompts abgeleitet. Konnten Nutzerantworten direkt als Prompt übernommen werden, wurden diese so belassen und nicht weiter korrigiert. So blieben Formulierungen wie *„Wie füge ich bei Word Hoch Zahlen ein, z.B Zwei zum Quadrat?"* und Rechtschreibfehler auch in der Angabe der Software wie *„Wie sicher ist Telegramm?"* bestehen. Der Grund dafür war, dass in dem Nutzungsszenario auch nicht ideal formulierte Prompts vorkommen können. ChatGPT müsste also in der vorgesehenen Nutzung als Software Erweiterung auch solche Prompts zufriedenstellend beantworten.

Nachdem die Daten gefiltert und die Prompts geschrieben waren, wurden sie nach Software-Kategorie und Software sortiert. Dies diente nicht nur der besseren Übersicht, sondern auch der Vereinheitlichung der Benennungen der Software. Jedem Prompt, der den Namen der Software noch nicht enthielt, wurde dieser hinzugefügt. Das ist notwendig, da ChatGPT noch nicht in der Software eingebunden ist und somit den entsprechenden Bezug mit übergeben bekommen muss.

4.2 Antwortengenerierung

Die Generierung der Antworten erfolgte ab Mitte November 2023. Es war vorgesehen, hierfür das ChatGPT-4 Modell zu nutzen. Allerdings wurde dann genau zu der Zeit die Anmeldung zu ChatGPT Plus pausiert, sodass zunächst auf ChatGPT-3.5 zurückgegriffen werden musste.

Wie im Anwendungsszenario in Unterabschnitt 3.1.1 beschrieben, ist das Ziel ChatGPT in die Software Systeme einzubinden. Der Platz zur Beschreibung der Erklärungen ist daher begrenzt. Entsprechend sollten die Antworten von ChatGPT kurz aber präzise ausfallen. Ferner werden die Erklärungen in einer Nutzerstudie ausgewertet. Dafür ist ein wichtiges Kriterium, dass die generierten Inhalte für den Nutzer schnell zu lesen und zu bewerten sind. Das ist auch im Anwendungsszenario selbst wünschenswert.

Um dies zu erreichen, wurden verschiedene Herangehensweisen getestet. Die Vorgaben an die Antworten von ChatGPT wurden hierbei anstelle der Anpassungsoptionen im Chat selbst vorgenommen. Es wurde sich aktiv gegen das Erstellen eines bzw. mehrerer eigener ChatGPT-Modelle entschieden. Im Szenario ist vorgesehen, ChatGPT ohne vorherige Anpassung einzubinden. Eine Individualisierung auf die jeweilige Software ist zudem bei über 70 verschiedenen Software-Systemen nicht im zeitlichen Rahmen dieser Arbeit darstellbar.

Zunächst wurden die Prompts einzeln eingegeben, mit dem Zusatz am Anfang des Chats, möglichst kurz zu antworten. Aufgrund der genutzten freien Version von ChatGPT war die Anzahl Anfragen pro Stunde begrenzt. Außerdem waren die Antworten zunächst sehr kurz und wurden dann mit zunehmender Anzahl Anfragen wieder deutlich länger, mit der Tendenz, sehr ausführlich zu sein. Auch wurden Schritt-für-Schritt Anleitungen generiert, welche eine Einbindung in die Studie aufgrund ihres Umfangs erschweren würden. Im Anwendungsfall wäre vor allem bei Apps hierfür in der Form auch kein Platz.

Aus diesen Gründen wurden mehrere Prompts als Liste in einer Anfrage mit dem Zusatz gebündelt, eine Tabelle mit den Antworten zu generieren. Hierdurch wurde automatisch auch die Länge der Antworten relativ einheitlich begrenzt. Die Angabe, eine angemessene oder sinnvolle Länge zu wählen, spielte dabei im Vergleich dazu, sie wegzulassen, keine signifikante Rolle. Dagegen fielen die Antworten mit dem Zusatz „kurz und einfach zu erklären“ deutlich kürzer aus.

Die kurzen Antworten fassten die längeren Antworten teils gut zusammen; die Antwort enthielten die gleichen oder wichtigsten Informationen in weniger Text. Diese Art Antwort wäre in den meisten Fällen anstrebenswert:

Kurz und einfach: *„Schreibe die Basiszahl, markiere sie und dann geh zu "Schriftart" und wähle das Hochstelltaste-Symbol (x^2) aus.“*

Sinnige Länge: *„Für Hochzahlen in Word markiere den Text, der hochgestellt werden soll, gehe zu "Start" > "Schriftart" und aktiviere "Hochgestellt". Alternativ verwende die Formel-Funktion unter "Einfügen" > "Gleichung", um das gewünschte mathematische Symbol einzufügen.“*

Beispielhafte Antworten von ChatGPT-3.5 zu der Software *Word*

Allerdings führte die Verknappung teilweise auch zu unzufriedenstellenden Antworten, wenn die Frage nicht präzise gestellt wurde. Beispielsweise wurden Nachfragen, ob es möglich ist, etwas zu tun, mit ja oder nein beantwortet ohne auszuführen, wie:

„Ja, in der Youtube-App kannst du Wiedergabelisten erstellen und abspielen.“

Daher wurde entschieden, mit den Antworten der Tabelle mit der Angabe „in sinniger Länge zu antworten“ zu arbeiten.

Die Qualität der damit generierten Antworten fiel subjektiv gesehen recht unterschiedlich aus. Innerhalb einer Anfragesitzung erschienen die generierten Antworten relativ konstant auf einem ähnlichen Niveau in Bezug auf die Länge und Formulierung der Antworten. In einer Session wurden die Fragen häufiger direkt, kurz und knapp beantwortet während in einer anderen die Antworten vermehrt eine Zusammenfassung der Frage enthielten. Diese Beobachtungen wurden jedoch nicht genauer untersucht oder quantifiziert, da nicht dokumentiert wurde, in welcher Session welche Anfragen gestellt wurden. Gleiche Wahrnehmung galt unabhängig der verwendeten Version von ChatGPT sowie beim Wechsel des Chats.

Im Januar des Jahres 2024 war es dann möglich, ChatGPT plus zu nutzen. Zu allen Prompts wurden mit ChatGPT-4 erneut Antworten generiert. Zu jedem Prompt wurden zwei Antworten aus unterschiedlichen Chats erzeugt. Es wurde wie zuvor, folgende Anweisung vor der Auflistung der Prompts mitgegeben:

„Alle Fragen bitte in einer sinnigen Länge beantworten und als tabelle ausgeben mit: Frage | Antwort“

4.3 Vorbereitungen zur Studie

Zur Vorbereitung der Daten auf die Studie wurden die Datensätze erneut in ein einheitliches Format gebracht und nicht benötigte Einträge, wie die Oberkategorien der Software sowie Leerzeilen, entfernt. Außerdem wurden die Prompts nach der Art der Frage bzw. entsprechend der Erklärung kategorisiert in Wie-, Was-, und Warum-Fragen, um später die kategoriespezifischen Nutzerbewertungen einbinden zu können. Soweit möglich geschah die Einordnung automatisch anhand des Anfangs des Prompts. Anschließend wurde die Kategorisierung manuell geprüft und ergänzt. Wo-Fragen wurden als Wie-Fragen eingeordnet, da diese meist darauf ausgelegt waren, wo und somit wie eine bestimmter Sachverhalt zu finden ist. Ähnlich wurden Prompts behandelt, welche nach der Möglichkeit, etwas Bestimmtes zu tun, fragten.

Des Weiteren wurden die Anzahl der Prompts für die Studie reduziert, um mehrere Bewertungen zu einer Antwort von ChatGPT zu erhalten und damit die Generalisierbarkeit zu erhöhen.

Von den insgesamt 227 entwickelten Anfragen wurde zufällig ein Fragenpool von 75 Fragen mit je zwei Antworten von ChatGPT-4 erstellt. Jeder Teilnehmer bekommt daraus 30 zufällig ausgewählte Erklärungen, zu denen er seine Bewertungen abgibt. Bei einer geplanten Teilnehmerzahl von 50 erhält somit jede Antwort von ChatGPT potenziell 10 Bewertungen von verschiedenen Teilnehmern:

$$10 Bewertungen_{proAntwort} = \frac{30 FA_{Paare} * 50 Teilnehmer}{75 FA_{gesamt} * 2 FA_{Varianten}}$$

Die finalen Daten enthielten die Angabe zur Software, die Kategorie, die Markierung der zur Studie ausgelosten verwendeten Einträge, den Prompt und die beiden Antworten von ChatGPT-4 zum jeweiligen Prompt.

Kapitel 5

Nutzerstudie I: Online-Studie

Zur Erfassung der subjektiven Metriken, welche in Abschnitt 3.2 beschrieben sind, wurde eine quantitative Nutzerumfrage durchgeführt. In dem nun folgenden Kapitel werden Vorgehen und Ergebnisse erläutert.

5.1 Methodik der Online-Studie

Um die zuvor aufgestellten subjektiven Metriken wie Zufriedenheit und Vertrauen quantitativ auswerten zu können, werden in einer Online-Studie Nutzerbewertungen zu den Erklärungen von ChatGPT gesammelt. In Abbildung 5.1 ist das Vorgehen dargestellt. Zunächst wird der Aufbau der Weboberfläche entwickelt, welche in erster Linie die Anfragen und Erklärungen sowie Bewertungsoptionen zeigt. Anschließend wird auf dieser Basis die Studie programmiert und Datenbanken zur Speicherung der Nutzerantworten angelegt. Nach Fertigstellung wurde die Online-Studie freigeschaltet und verteilt. Abschließend wurden die Ergebnisse gesammelt und die Daten analysiert.

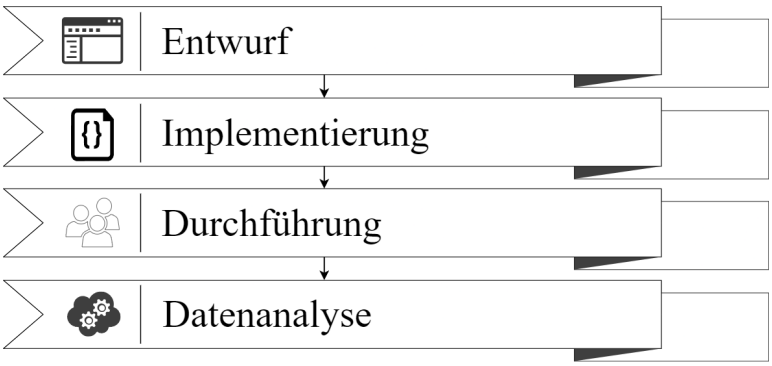


Abbildung 5.1: Ablauf der Online-Studie

5.1.1 Entwurf

Die entwickelte Webanwendung diente rein zur Durchführung der Studie. Daher wurden keine weiterführenden Tests wie zum Beispiel zur Usability [42] durchgeführt. Die Idee zur Gestaltung der Oberfläche war eine möglichst klare und einfache Struktur zu erstellen, welche angelehnt an die Nutzung von ChatGPT ist.

Wichtig war die Bereitstellung der Frage bzw. des Prompts, die zugehörige Antwort von ChatGPT sowie die vom Teilnehmer zu bewertenden Kriterien. Unter dem eingegebenen Alias des Nutzers wird zusätzlich der Name der Software angezeigt, auf welche sich der Prompt bezieht. Um eine Übersicht über den aktuellen Fortschritt in der Umfrage zu erhalten, wird ein Fortschrittsbalken angezeigt. Nachdem die Oberfläche grob entworfen wurde, wurde diese umgesetzt. Die gewählten Farben entsprechen dem Logo von ChatGPT sowie der Chat-Darstellung im Darkmode. Ein Ausschnitt der Hauptfrageseite ist in Abbildung 5.2 dargestellt.

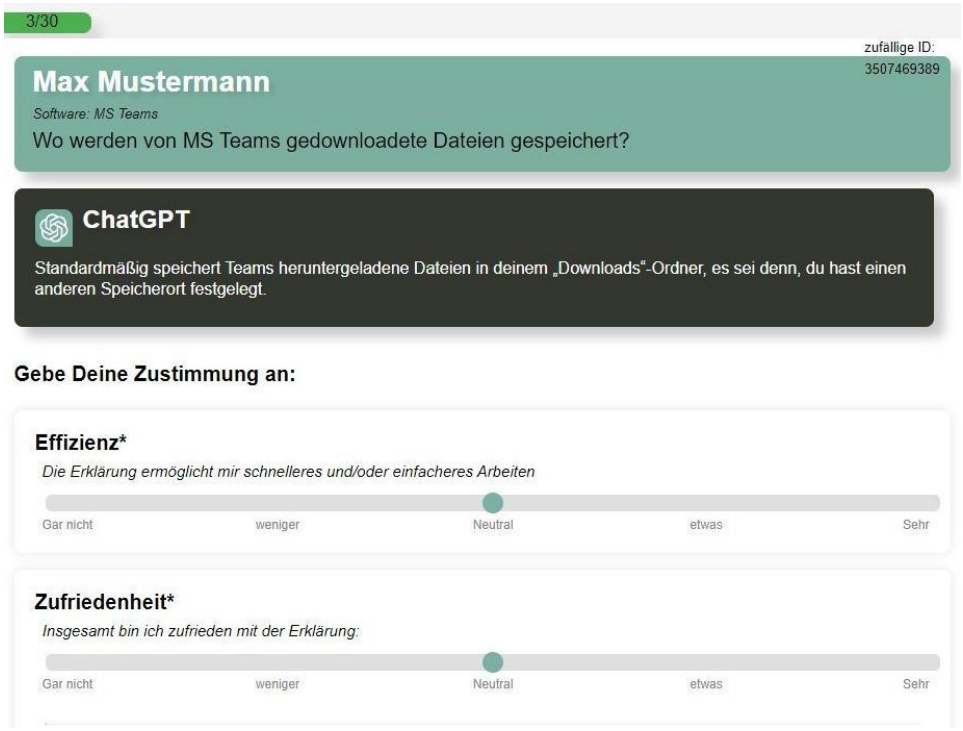


Abbildung 5.2: Bewertung der Zustimmung: Ausschnitt einer Frage-Antwort-Seite der Online-Studie

Jedem Teilnehmer wird eine zufällige ID zugewiesen, um später die Daten pseudonymisiert verarbeiten zu können.

5.1.2 Implementierung

Die Umfrage lief über den Linuxserver der Universität mittels Docker. Es wurde zunächst die Dockerumgebung aufgesetzt, sodass mittels docker-compose der Container alle benötigten Images startet:

- MongoDB: Die verwendete Datenbank zur Speicherung der gesammelten Daten sowie zum Abrufen der zuvor eingespeicherten Prompts und Antworten

- Mongo-Express: Als grafische Oberfläche zur erleichterten Bedienung der Datenbank

- Survey-App: Die mit node.js und express.js für die Studie programmierte Anwendung. Die Hauptanwendung besteht aus einer HTML-Datei unter Verwendung von CSS und JavaScript

- Nginx: Für das Weiterleiten der HTTP-Anfrage der Server URL (platon.se.uni-hannover.de) auf die lokal verwendete IP bzw. Adresse (survey-app:3000)

Des Weiteren wurde ein Verzeichnis (Volume) für die persistente Speicherung der Daten der Datenbank angelegt. Über den Docker Hub wurde das Projekt dann auf den Server übertragen und so angepasst, dass es lauffähig war. Das Volume konnte auf diesem Wege nicht mit übertragen werden, sodass die Datenbank manuell mit den Werten erneut bestückt wurde.

Im Zuge der Programmierung wurde ChatGPT zum Debugging sowie nach Anpassungsmöglichkeiten der CSS befragt, welche dann einen Ansatz zur Gestaltung des optischen Aussehens lieferten.

5.1.3 Durchführung

Die Studie wurde per Link online bereitgestellt und aktiv über einen Zeitraum von zwei Wochen ab Mitte Februar 2024 verteilt und bis zur Datenauswertung bis Ende März online gelassen. Die Umfrage wurde an Bekannte und Verwandte verteilt, mit der Bitte, die Studie weiterzuverteilen. Außerdem wurden Mitstudenten angesprochen und ein online Pinnwand Post im Universitätsportal erstellt.

Da es das Ziel ist, herauszufinden, ob die Antworten von ChatGPT unabhängig der Person zufriedenstellend sind, wurde ein möglichst breit gefächertes Teilnehmerspektrum angestrebt. Es wurden keine Vorkenntnisse benötigt; das Hineindenken in die Software sowie die Fragen und Antworten wurde dadurch jedoch erleichert.

Aufbau der Studie

Der Fragebogen umfasst neben dem Hauptteil mit den Fragen zur Bewertung der ChatGPT Erklärungen zu den Prompts noch eine Einführung zur Studie, eine Frageseite über den Nutzer sowie eine Abschlussfrageseite (Anhang A.1).

In der Einführung wird die Studie kurz vorgestellt und der Nutzer muss bestätigen, dass er über 18 Jahre alt ist. Außerdem muss der Nutzer vor dem Beginn der Studie zustimmen, dass seine Daten pseudonymisiert verarbeitet und veröffentlicht werden dürfen.

Auf der folgenden Seite werden Daten zum Nutzer erhoben; er kann ein Alias vergeben, welches dann über den Prompts auf den nächsten Seiten angezeigt wird. Außerdem kann er Alter, Geschlecht sowie seinen aktuellen Arbeitsstatus (studierend, arbeitend) angeben. Abschließend wird nach den aktuellen Erfahrungen mit ChatGPT gefragt. Neben der allgemeinen Einstellung des Nutzers gegenüber ChatGPT wird sein Vertrauen in die gelieferten Antworten abgefragt. Außerdem war anzugeben, wie oft der Teilnehmer ChatGPT benutzt. In einem Kommentarfeld können zusätzliche Anmerkungen gegeben werden.

Darauf folgen die 30 Prompts mit einer Erklärung von ChatGPT. Der Nutzer hat bis zu insgesamt 15-mal die Möglichkeit, sich ein anderes Beispiel geben zu lassen, sollte er mit der erklärten Software gar nichts anfangen können. Unter den zu bewertenden Antworten von ChatGPT folgen die in Abschnitt 3.3 erarbeiteten Fragen. Zunächst erfolgt die kategoriespezifische Frage; abhängig davon, ob es eine Wie-Frage ist, wird hier die wahrgenommene Effizienz(-steigerung) und ansonsten die Angemessenheit abgefragt. Darauf folgen die Fragen bezüglich der Zufriedenheit, Verständlichkeit und wahrgenommenen Effektivität des Nutzers. Diese drei Fragen sind auf jeder Seite die gleichen, sodass der Teilnehmer die Möglichkeit hat, schnell und intuitiv zu antworten. Zum Abschluss jeder Seite hat der Teilnehmer die Möglichkeit, einen Kommentar zu verfassen. Das Kommentarfeld ist dazu da, dass der Nutzer alles was ihm auffällt noch anmerken zu können. Das kann zum Beispiel sein, dass eine Frage nicht korrekt beantwortet wurde oder der Nutzer etwas nicht verstanden hat.

Am Ende der Studie wird der Nutzer danach gefragt, wie zufrieden er insgesamt mit den Antworten von ChatGPT war. Außerdem soll er beurteilen, inwieweit die Erklärungen zur Verständlichkeit bzw. Transparenz der jeweiligen Software beigetragen haben. Der Teilnehmer wird erneut nach seinem Vertrauen in die Antworten von ChatGPT gefragt. Ferner soll er reflektieren, für wie hilfreich er die Erklärungen insgesamt im Nachhinein

befindet. Abschließend wird festgehalten, wie sehr der Teilnehmer sich wünschen würde, dass ChatGPT im Software-System selbst zur Klärung von Fragen zur Verfügung steht. Es ist zu klären, ob der Teilnehmer einen Mehrwert in der eigentlichen Software-Integration sieht. Auch auf dieser Seite steht am Ende ein Kommentarfeld zur Verfügung.

Zum Schluss wird dem Nutzer für seine Teilnahme gedankt und er darüber informiert, dass seine Daten erfolgreich gespeichert wurden.

Erfassung der Daten

Die Bewertungen der Nutzer zu den Erklärungen von ChatGPT werden mithilfe von Likert-Skalen erfasst und können somit als Zahlenwert gespeichert werden. Diese Werte werden besonders oft abgefragt. Daher wurden hier 5-stufige Likert-Skalen (mit Werten im Bereich von 1 bis 5) gewählt, um dem Teilnehmer mehrere Wahlmöglichkeiten zur Verfügung zu stellen und ihn gleichzeitig nicht zu überfordern. Die erste Bewertung des Vertrauens sowie die Fragen der Abschlussseite bestanden aus 7-stufigen Likert-Skalen (mit Werten im Bereich von 1 bis 7), um detaillierteres Feedback zu erhalten. Jede Seite erzeugt dabei einen eigenen Datenbankeintrag, welcher über die zufällig generierte Nutzer ID pseudonymisiert zugeordnet werden kann. Die Angaben zu Geschlecht und Arbeitsstatus werden mittels Single-Choice Felder abgefragt. Hier gibt es jeweils die Möglichkeit ‚Keine Angabe' bzw. ‚Anderes' auszuwählen. Optionale Kommentare zu allen Seiten werden in Textfeldern erfasst.

5.1.4 Datenanalyse

Die Analyse der in der Online-Studie gesammelten Daten wurde mittels eines Jupyter Notebooks durchgeführt. Die Daten wurden zunächst gefiltert und auf dieser Basis die Auswertung der Angaben zum Nutzer sowie die Bewertungen der Erklärungen durchgeführt. Unabhängig von der Filterung wurden alle gesammelten Kommentare betrachtet und ausgewertet.

Für die statistische Auswertung wurden nur die Bewertungen der Teilnehmer mit einbezogen, welche den Fragebogen komplett bis zum Ende bearbeitet hatten. Außerdem wurden die Daten manuell überprüft: Neben dem Durchsehen der Kommentare wurde nach Mustern gesucht, die auf ein zufälliges, gezieltes oder komplett einheitliches Bewerten der Fragen hindeuteten. Hier wurden drei Nutzer aussortiert. Einer davon hatte keinen der Schieberegler der Likert-Skalen verändert, während die anderen beiden angaben, die Umfrage nicht verstanden zu haben. Sonstige herausstechende Daten waren bereits durch das Nichtbeenden der Umfrage automatisch herausgefiltert. Andere Auffälligkeiten konnten in den Daten nicht entdeckt werden.

Vorgehen zur Auswertung der Daten

Zunächst wurden die demografischen Angaben der Teilnehmer ausgewertet. Des Weiteren erfolgte die Auswertung der allgemeinen Einstellung des Teilnehmers sowie der Angaben zur Erfahrungen und Verwendungshäufigkeit von ChatGPT. Außerdem wurde das anfängliche Vertrauen des Teilnehmers mit dem am Ende der Umfrage angegebenen verglichen.

Als nächstes wurden die Verwertungen der Erklärungen analysiert. Hier wurde zunächst eine Gesamtauswertung über alle erhaltenen Bewertungen durchgeführt, um einen Überblick über die generelle Güte der Erklärungen zu erhalten (RQ1, RQ2). Die Bewertungen wurden außerdem auf Unterschiede in der Antwortengebung zwischen den Geschlechtern männlich und weiblich sowie verschiedenen Altersstufen (bis 40, ab 40 und ab 60) untersucht. Zu jedem Prompt aus dem Fragenpool wurden zwei Antworten von ChatGPT generiert. Auch hier wurde eine Übersicht über alle Bewertungen jeweils zu Antwort 1 und Antwort 2 erstellt, um diese vergleichen zu können (RQ4). Anschließend wurde für jeden einzelnen Prompt ein Diagramm erstellt, um Unterschiede bei einzelnen Fragen und zugehörigen Erklärungen zu ermitteln (RQ1, RQ2, RQ4). Bei jeder Frage wurde zusätzlich in den beiden Antwortgebungen von ChatGPT unterschieden (RQ4). Weiter wurden die Daten in die Kategorien der Erklärungstypen und der Art des Erklärungsbedarfs unterteilt (RQ1, RQ4). Für beide Kategorien wurde ebenfalls auf Unterschiede zwischen den generierten Erklärungen des ersten (Antwort 1) und zweiten Chats (Antwort 2) geprüft (RQ4). Abschließend wurde eine Auswertung der Antworten für jeden Teilnehmer vorgenommen, um individuelle Unterschiede in den Bewertungen zu identifizieren, und die gegebenen Kommentare betrachtet (RQ3).

Die Bewertung der Abschlusseite beinhaltet die Aspekte Gesamtzufriedenheit, Steigerung der Software Transparenz, Effektivität insgesamt sowie den Wunsch nach einer Integration von ChatGPT in Software Systeme (RQ1, RQ2). Auch hier wurde auf Unterschiede zwischen je dem Geschlecht und den Altersgruppen geprüft. Die Kommentare zu Beginn sowie zum Abschluss der Umfrage werden analysiert.

Teilnehmer

Insgesamt beendeten 70 Personen die Umfrage und 2241 Datenbankeinträge zu den Frage-Antwort-Seiten wurden gesammelt. Ausgewertet wurden 67 Teilnehmer und 2004 Daten zu den Frage-Antwort-Paaren. Bei der Übertragung der Daten gingen offensichtlich sechs Datenpunkte verloren. Da die betroffenen Teilnehmer den Fragebogen jedoch beendet hatten und durch den Zufallsfaktor ohnehin jede Erklärung eine unterschiedliche Anzahl

Bewertungen erhielt, wurden die erhaltenen Antworten dieser Teilnehmer in der Auswertung mit berücksichtigt. Alle Werte wurden wenn nicht anders angegeben auf zwei Nachkommastellen gerundet.

Von den Teilnehmern waren 32 männlich und 31 weiblich, 1 Person divers und 3 ohne Angabe. Das Alter lag zwischen 18 und 71 Jahren, wobei 4 Personen ihr Alter nicht angaben. Das mittlere Alter betrug etwa 33 Jahre (SD = 15,47), der Median lag bei 26 Jahren. 25 der Befragten gaben an, aktuell zu arbeiten, 13 zu studieren und 21 zu arbeiten und zu studieren.

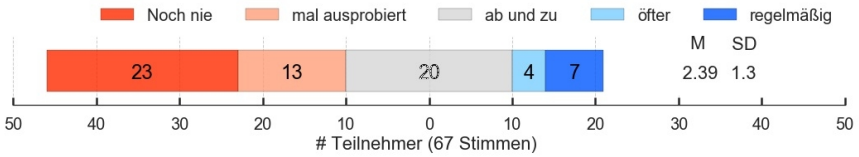


Abbildung 5.3: Erfahrungen mit ChatGPT

Die Mehrheit der Nutzer verwenden ChatGPT ab und zu oder seltener, davon 23 noch nie (Abbildung 5.3). 7 Teilnehmer verwenden ChatGPT regelmäßig.

Die allgemeine Einstellung der Nutzer gegenüber ChatGPT gaben die meisten als neutral (28 Probanden) bis eher positiv (24 Probanden) an. Der Mittelwert lag bei 3,16; der Wert 3 entspricht dabei *‚neutral‘*. Das Vertrauen der Teilnehmer in die Antworten von ChatGPT war zu Beginn der Umfrage im Durchschnitt mit 3,7 etwas geringer als *‚neutral‘* (Wert von 4) und zum Ende der Umfrage im Durchschnitt fast *‚neutral‘* (Abbildung 5.4).

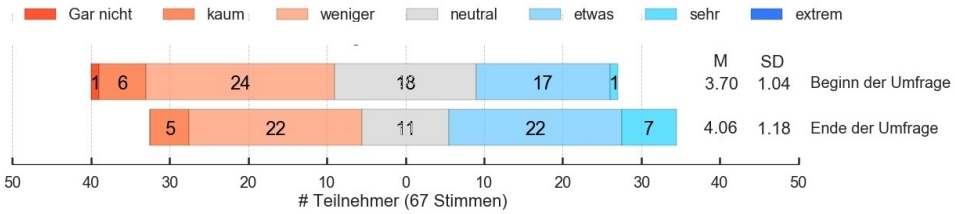


Abbildung 5.4: Höhe des Vertrauens in die Antworten von ChatGPT

Kommentare zu Beginn der Umfrage Von allen Teilnehmern, die die Umfrage gestartet hatten, gaben Fünf zu Beginn einen Kommentar. So sei ChatGPT für einen *„ersten Input sehr hilfreich“* und würde öfter zum Formulieren genutzt. Wissenschaftliche Fakten würden geprüft. Ein Nutzer merkte an, dass seiner bisherigen Erfahrung nach *„[a]lles was ChatGPT bisher leisten sollte [...] falsch [war]“*.

Ein Anderer hat *„[V]ertrauen in Chat GPT“* und ist *„grundsätzlich eher positiv, unter der Voraussetzung, dass die anwendende Person die Dinge, die*

Chat GPT ausgibt, auch hinterfragt und anzuwenden weiß und ein gewisses Hintergrundwissen bereits hat“. Weiter merkte ein Nutzer an, dass sofern „*man die Frage richtig formuliert hat*“ ChatGPT als Lernhilfe und für das Generieren von Codebeispielen nutzen kann.

5.2 Ergebnisse der Online-Studie

In diesem Teil werden die mit Jupyter Notebook analysierten Daten der Frage-Antwort-Seiten sowie der Abschlusseite des Fragebogens dargelegt.

Für die Bewertungen der Erklärungen wurden mehrere Ansätze ausgewertet (Unterabschnitt 5.1.4), jeweils über die einzelnen zu bewertenden Aspekte der kategoriespezifischen Frage (Effizienz oder Angemessenheit), Zufriedenheit, Verständlichkeit und Effektivität. Da 5-stufige Likert-Skalen verwendet wurden entspricht der neutrale Mittelwert dem Wert 3.

Die Abschlussseite hält die zusammenfassende Meinung des Teilnehmers fest und wird abschließend gesondert ausgewertet. Bei den dort gestellten Fragen wurden 7-stufige Likert-Skalen verendet, wodurch das Neutrum hier dem Wert 4 entspricht. Alle Werte wurden, wenn nicht anders angegeben, auf zwei Nachkommastellen gerundet.

5.2.1 Bewertung der Erklärungen: Frage-Antwort-Paare

Die Bewertungen über alle Teilnehmer und alle Erklärungen hinweg ergaben über 2000 Datensätze (hier: Stimmen). Eine Stimme entspricht dabei der Bewertungen eines Frage-Antwort-Paars, also einem Prompt und der Erklärung.

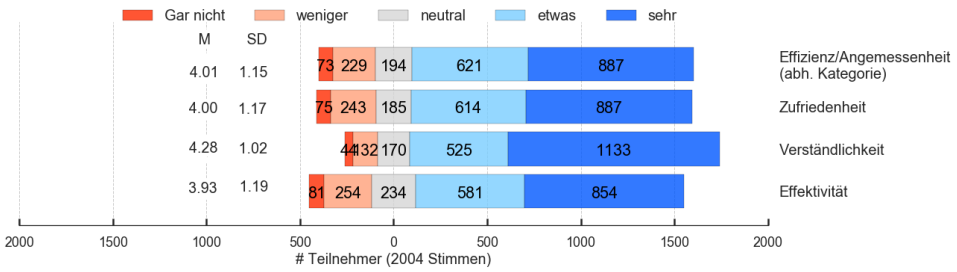


Abbildung 5.5: Bewertung über alle Erklärungen über alle Teilnehmer

Die meisten Stimmen der Teilnehmer drückten in allen zu bewertenden Aspekten vorwiegend *etwas* oder *sehr* große Zustimmung aus (Abbildung 5.5). Die Verständlichkeit wurde dabei am besten bewertet (M = 4,28). Dahinter lagen die Zufriedenheit und der kategoriespezische Aspekt mit einem Mittelwert um 4, welche *etwas* Zustimmung darstellt. Die Unterteilung der beiden Aspekte Angemessenheit und Effizienz ist im Anhang grafisch

widergegeben (Anhang C). Knapp 17 % der 2000 Stimmen drückten aus, dass die Erklärungen für den Teilnehmer *weniger* (254 Stimmen) bis *gar nicht* (81 Stimmen) hilfreich waren. Insgesamt wurde die Effektivität zustimmend bewertet (M = 3.93).

Unterscheidung nach Alter und Geschlecht Vergleichend wurden die Antworten nach Alter und Geschlecht getrennt untersucht. Hierbei werden in allen Aspekten vergleichbare Ergebnisse erzielt (Tabelle 5.1).

Aspekt		**Geschlecht**		**Alter**		
		M	*W*	*Bis 40*	*Ab 40*	*Ab 60*
Kategoriespezifisch	M	3,98	4,02	4,01	3,92	3,97
	SD	1,15	1,17	1,15	1,19	1,15
Zufriedenheit	M	4,02	3,97	3,98	3,98	4,15
	SD	1,16	1,17	1,17	1,17	1,07
Verständlichkeit	M	4,28	4,27	4,25	4,29	4,31
	SD	1,01	1,03	1,02	1,04	0,98
Effektivität	M	3,98	3,96	3,92	3,92	4,03
	SD	1,21	1,18	1,20	1,20	1,09

Tabelle 5.1: Vergleich der Bewertungen nach Alter und Geschlecht

Unterscheidung beider generierten Antworten von ChatGPT Zu jedem Prompt wurden zwei Antworten von ChatGPT generiert, wovon eine zufällig für den Prompt bzw. Teilnehmer ausgewählt wurde. Die erzeugten Erklärungen *Antworten 1* erhielten 1012 Stimmen und *Antworten 2* die anderen 992 Stimmen. Auch hier wurden die Bewertung der Teilnehmer auf mögliche Unterschiede untersucht (Tabelle 5.2). Insgesamt wurden beide Antworten ähnlich bewertet.

Aspekt		**Antworten 1**	**Antworten 2**
Kategoriespezifisch	M	4,00	4,02
	SD	1,16	1,14
Zufriedenheit	M	3,98	4,01
	SD	1,20	1,13
Verständlichkeit	M	4,28	4,31
	SD	1,06	0,96
Effektivität	M	3,91	3,96
	SD	1,22	1,15

Tabelle 5.2: Vergleich der beiden generierten Antworten

Bewertungen pro Prompt

Die höchsten Zustimmungen in allen Aspekten erhielten die Erklärungen zu den Prompts 1, 8, 11, 43, 51 und 72 (Mittelwerte um M = 4,7 und höher, SD < 1, teilweise SD < 0,5). Dies entsprach den folgenden Anfragen:

- 1: *Wie füge ich bei Word Hoch Zahlen ein, z.B Zwei zum Quadrat?*
- 8: *Wie kann ich bei Teams weitere Kontakte zu einem Meeting einladen?*
- 11: *Wie kann ich bei MS Teams während eines Meetings die Chat Benachrichtigungen abschalten?*
- 43: *Sieht man, ob jemand von meiner Story bei Instagram einen Screenshot gemacht hat?*
- 51: *Wie kann ich bei Google Chrome die Standardsuchmaschine ändern?*
- 72: *Wie/Wo kann ich den Verlängerungskey bei Avira eingeben?*

Die Erklärungen mit der wenigsten Zustimmung waren die Prompts 24 und 65 (Abbildung 5.6). Der Mittelwert der Aspekte, abgesehen von der Verständlichkeit, lag in beiden Fällen unter der Bewertung *neutral*. Die entsprechenden Prompts dazu waren:

- 24: *Wie kann ich bei Eclipse die Hibernate-Anbindung richtig konfigurieren?*
- 65: *Was bedeuten die Kürzel in den Einstellungen bei Chirp?*

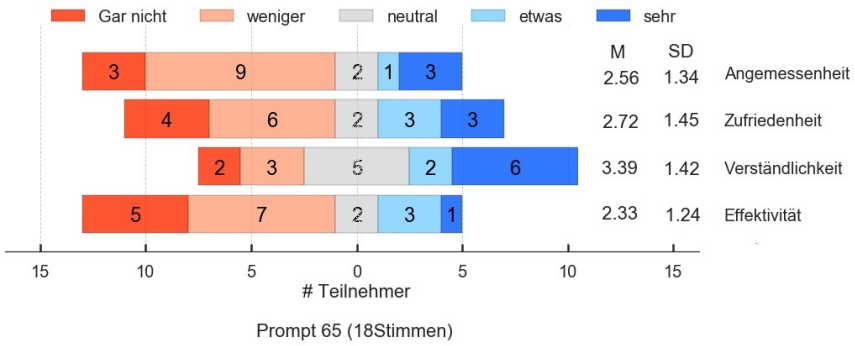


Abbildung 5.6: Erklärungen zu Prompt 65 mit der geringsten Zustimmung: *Was bedeuten die Kürzel in den Einstellungen bei Chirp?*

Unterscheidung beider generierten Antworten von ChatGPT Die beiden generierten Antworten des selben Prompts wurden überwiegend ähnlich bewertet. In den meisten Fällen unterschieden sich die Mittelwerte im Rahmen einer Zustimmungsstufe. Bei den Erklärungen zu einem Prompt (Prompt 32) war die Diskrepanz besonders hoch (Abbildung 5.7):

- 32: *Kommen Nachrichten bei Whatsapp im Nachhinein an, wenn eine Person mich entblockt?*
  - Antwort 1: *Ja, Nachrichten, die gesendet wurden, während du blockiert warst, werden jedoch nicht nachträglich zugestellt.*
  - Antwort 2: *Nachrichten, die während der Blockierung gesendet wurden, werden nicht zugestellt, aber nach der Entblockung gesendete Nachrichten kommen an.*

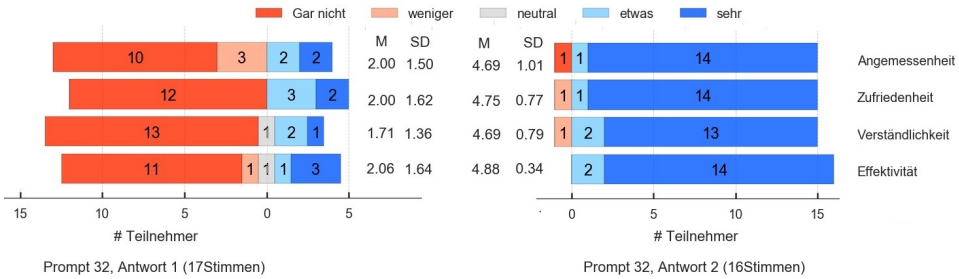


Abbildung 5.7: Bewertung der Erklärungen zu Prompt 32 Antwort 1 (links) und Antwort 2 (rechts)

Bewertungen nach Erklärungsbedarf Zu den Arten des Erklärungsbedarfs wurden folgende Anzahl Stimmen erhalten: *Interaktion* 956 Stimmen, *Systemverhalten* 599 Stimmen, *Security* 293 Stimmen und *Domainwissen* 156 Stimmen. Die Ergebnisse sind vergleichbar mit denen, die in der Gesamtbewertung erhalten wurden Abbildung 5.5. Die einzelnen Grafiken finden sich im Anhang (Anhang C).

Bei allen Arten erhielt die höchste Zustimmung der Aspekt der Verständlichkeit, gefolgt von dem kategoriespezifischen Aspekt (Angemessenheit oder Effizienz), der Zufriedenheit und der Effektivität (**??**).

Für die Anfragen, welche die *Interaktion* mit dem System betreffen, wurden die höchsten Zustimmungen zu den jeweiligen Aspekten angegeben. Über die Hälfte aller Bewertungen stimmten jeweils *sehr* zu, ein weiteres Viertel *etwas*. Bei den anderen Arten des Erklärungsbedarfs war die Anzahl der Zustimmung zwischen *etwas* und *sehr* ausgeglichener, mit Tendenz zu *sehr* großer Zustimmung.

Unterscheidung beider generierten Antworten von ChatGPT Die Betrachtung der Arten des Erklärungsbedarfs abhängig von ChatGPTs gegebenen *Antwort 1* oder *Antwort 2* ergab keine auffälligen Unterschiede.

Bewertungen nach Erklärungstyp Für den Erklärungstypen der *Wie-Erklärungen* wurden 1147 Stimmen gesammelt, bei den *Was-Erklärungen* 698 Stimmen und den *Warum-Erklärungen* 159 Stimmen. Die Ergebnisse sind vergleichbar mit denen, die in der Gesamtbewertung erhalten wurden Abbildung 5.5.

Die *Wie-Erklärungen* erhielten die höchsten Bewertungen in den einzelnen Aspekten. Zu jedem Aspekt drückten über 75 % der gesammelten Stimmen ihre Zustimmung aus. Davon waren an die 50 % *sehr* zustimmend. Die Mittelwerte lagen alle über 4; am höchsten bewertet wurde die Verständlichkeit (M = 4.36). Der kategoriespezifische Aspekt Effizienz wurde ähnlich wie die Zufriedenheit und Effektivität bewertet.

Bei den *Was-Erklärungen* ist das Verhältnis zwischen etwas und sehr zustimmenden Bewertungen fast ausgeglichen. Nur bei der Verständlichkeit gab es deutlich mehr *sehr* Zustimmende (M = 4,21). Die Bewertungen der Angemessenheit lagen leicht über denen der Effektivität und Zufriedenheit mit den generierten Antworten.

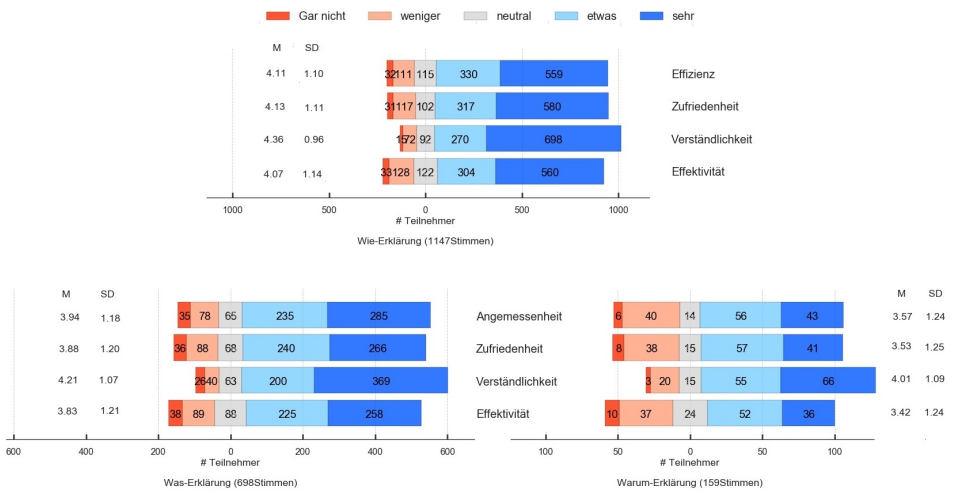


Abbildung 5.8: Zustimmung der Bewertungen nach Erklärungstyp, Wie-Erklärung (oben), Was-Erklärung (unten links), Warum-Erklärung (unten rechts)

Die generierten *Warum-Erklärungen* erhielten in den einzelnen Aspekten im Vergleich zu den anderen Erklärungstypen weniger Zustimmung Abbildung 5.8. Dabei wurden die Aspekte mit Ausnahme der Zufriedenheit (M = 4,01) ähnlich bewertet.

Unterscheidung beider generierten Antworten von ChatGPT Die Auswertungen für die beiden generierten Antworten von ChatGPT zu jedem Erklärungstypen unterschieden sich in ihren Mittelwerten kaum. Die größte Differenz war bei den *Antworten 1* und *Antworten 2* der *Warum-Erklärungen* zu verzeichnen. Hier unterschied sich der Mittelwert der einzelnen Aspekte zwischen 0,28 und 0,43.

Bewertungen nach Teilnehmer

Die Bewertungen der einzelnen Teilnehmer sind individuell. Teils gab es Nutzer, welche insgesamt und über alle Aspekte sehr zustimmend antworteten. Bei anderen fiel die Bewertung abhängiger vom jeweiligen Aspekt aus. Beispielsweise beurteilte ein Nutzer die Effektivität im Mittel über alle bewertete Erklärungen als neutral, die Verständlichkeit zumeist hoch (M = 4,43). Dies spiegelte sich auch in seiner Abschlussbewertung wider, in der er die Effektivität als weniger hilfreich beurteilte. Ähnlich urteilte ein andere Nutzer. Ein Weiterer urteilte über alle Bewertungen im Mittel alle Aspekte nahe neutral (je SD ≈ 1). Diese Art Nutzer waren seltener vertreten. Insgesamt bewerteten die meisten Nutzer alle Aspekte mit hoher Zustimmung.

Kommentare der Teilnehmer Insgesamt kommentierten 26 Nutzer im Verlauf der Umfrage bei den Bewertungen der Erklärungen von ChatGPT. Von 6 Teilnehmern gab es positive Äußerungen, wenn etwas gut beantwortet wurde. Im Gegensatz dazu merkten mehrere Nutzer bei Erklärungen an, dass etwas keinen Sinn ergab (5 Teilnehmer), die Frage nicht richtig (2 Teilnehmer) oder nur teilweise beantwortet wurde (5 Teilnehmer). Dabei wurde bemängelt, dass die Angaben zu allgemein, sehr vereinfacht und oberflächlich (über 15 Teilnehmer) bzw. zu kurz waren (8 Teilnehmer). Ein Nutzer schrieb beispielsweise, er würde *„gern noch wissen wie es genau geht“*.

Stellenweise seien die Erklärungen uneindeutig oder, wie im Fall von Prompt 32 (Abbildung 5.7), widersprüchlich (6 Teilnehmer). Insgesamt gaben mindestens 8 Nutzer kritische Kommentare zu ChatGPT:

- *„Für insider ok, für Leute ohne vorkenntnisse mangelhaft“*
- *„Chat gpt kann nicht antworten.“*
- *„Bei einer solchen Fragestellung erwarte ich vor und Nachteile. Erheblich ist wenig aussagend“*
- *„Es fehlt der Werbeaspekt“* oder *„Hört sich wie Werbung einer Bank an“*
- *„Verweis auf Dokumentation (Link) wäre wünschenswert.“*
- *„Es fehlen Beispiele“*

Ein Nutzer schrieb explizit, er hätte ChatGPT gern selbst ausprobiert:

> *„Für alle solche Fragen gilt: Das würde ich gerne einmal nach der Anleitung von [ChatGPT] probieren, um herauszufinden, ob die Anleitung wirklich funktioniert. Die Möglichkeit habe ich jedoch gerade nicht.“*

Ein weiterer Nutzer merkte an, er habe das Gefühl, dass ChatGPT sich von Suggestivfragen beeinflussen lassen würde. Er nannte das Beispiel:

> *„Wenn gefragt worden wäre ob es an zu neuen android Versionen liegen würde wäre die Antwort dass die neuen Androidversionen das Problem sind“*

Über 7 Benutzer äußerten, dass sie das Thema oder die Software, auf die die Frage Bezug nahm, nicht kennen würden. Einige schrieben weiter, dass ihnen daher die Beurteilung schwer falle oder nicht möglich sei.

5.2.2 Gesamtbewertung: Abschlussseite

Insgesamt waren die Teilnehmer *etwas* bis *sehr* zufrieden mit den generierten Erklärungen von ChatGPT (Abbildung 5.9). Keiner der Befragten gab an, dass er *gar nicht* oder *extrem* zufrieden gewesen sei. Die meisten Nutzer bewerteten außerdem, dass ChatGPT ihnen die im Prompt erwähnte Software verständlicher gemacht habe und die Erklärungen insgesamt hilfreich waren.

Die Meinung, sich ChatGPT als direkte Einbindung in der Software zu wünschen, war geteilt. Hier bräuchten 21 der 67 Teilnehmer nicht unbedingt eine Software Integration, davon 5 *gar nicht*. Entgegen dem wünschten sich 34 Nutzer ChatGPT als Hilfe im Software System, zwei davon *extrem*.

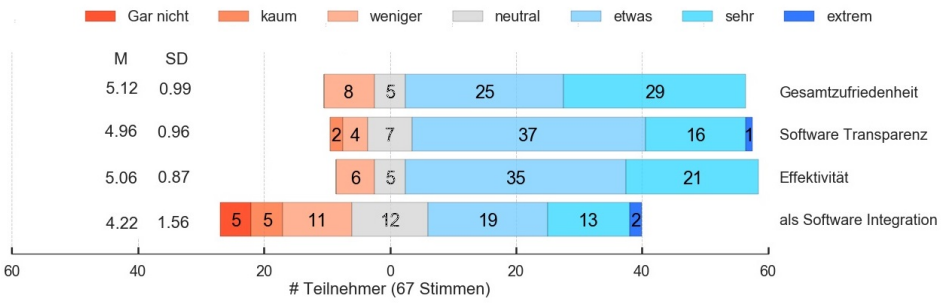


Abbildung 5.9: Abschlussbewertungen zu ChatGPT

Unterscheidung nach Alter und Geschlecht Die Untersuchung nach Alter und Geschlecht getrennt ergaben erneut kaum Abweichungen. Einzig die Software Integration wünschte sich die Altersgruppe ab 60 im Vergleich zum Durchschnitt weniger (M = 3.86).

Kommentare am Ende der Umfrage Positiv hervorgehoben wird, dass ChatGPT generell sehr hilfreich sein könne, um *„ein erstes Verständnis zu bekommen“* und *„einen Überblick über die Software zu gewinnen“*. Ein Nutzer gab an ChatGPT als Software Integration nutzen zu wollen unter der Bedingung, dass *„in sachen vollständigkeit und verständlichkeit laufend angepasst würde“*.

Weiter sei ChatGPT insbesondere hilfreich bei *„bestimmten Klickanleitung[en]“*, dem Generieren von Code und dem Weisen in die *„richtige Richtung“*. Ein Nutzer hob hervor, dass ChatGPT sich in *„allgemeinverständlicher Sprache“* ausdrücke.

Bedenken hatten mehrere Nutzer hinsichtlich der Seriosität und Herkunft der Quellen und Angaben von ChatGPT. Dementsprechend wären Nutzer *„mit dem Vertrauen zurückhaltender“*, würden sich *„nicht darauf verlassen“* und müssten *„alle antworten prüfen“*. Ein Nutzer sähe es daher *„als Risiko, [ChatGPT] in Software-System[en] integriert zu nutzen“*. *„Viele System[e] [würden ChatGPT] integrieren [...] obwohl dieses nicht ausgereift ist“*.

Allgemein bezeichnete ein Nutzer ChatGPT als *„Ausbaufähig“*, ein Weiterer bemängelte, dass *„die Antworten nicht genug in die Tiefe“* gingen. Eine Aussage deklarierte, dass teilweise Dinge erklärt würden, die *„schlicht falsch sind“*. Die gegebenen Antworten von ChatGPT in der Online-Umfrage seien abgesehen von enger gefassten Fragen *„zu allgemein, um eine direkte Hilfe darzustellen“*

Insgesamt befanden mehrere Nutzer, dass die Zufriedenheit mit den Antworten von ChatGPT von der gestellten Frage abhinge: *„Je genauer die Fragen gestellt werden, desto besser die Antworten“*; *„wenn man spezifischer nachfragen würde, würde es eventuell bessere Antworten geben“*. Außerdem hinge die *„Zufriedenheit und vor allem das Vertrauen [...] von der Korrektheit der Antworten ab“*. Die Korrektheit könne er jedoch nicht bewerten. Ähnliches wurde von einem anderen Nutzer angemerkt. Dieser meinte weiter, dass die *„Software Integration neutral [wäre], da es sehr auf das konkrete System ankomm[en würde]“*.

Es sei *„durchaus vorstellbar, dass ChatGPT in ein paar Jahren [...] leistungsfähiger ist und dessen Anwendungsbereiche dadurch vielseitiger sein werden“*. Dabei merkte ein Nutzer an, dass *„ChatGPT [...] sinnvoll und wichtig [sei], aber [...] reguliert werden [müsste] [...]“*. ChatGPT stünde *„erst am Anfang [und] ob es der Gesellschaft hilft[, wird] die Zukunft zeigen“*.

Kapitel 6

Nutzerstudie II: Interview-Studie

In diesem Kapitel wird die ergänzend zur Online-Studie durchgeführte qualitative Interview-Studie beschrieben. In dieser wurde die ChatGPT-Nutzung unter realen Bedingungen betrachtet und bewertet, um die vorherigen Ergebnisse der quantitativen Studie zu validieren und zu ergänzen.

6.1 Methodik der Interview-Studie

In der Interview-Studie hatte der Nutzer die Möglichkeit, selbst mit ChatGPT zu interagieren und eigene Fragen und Nachfragen zu stellen. Die Nutzer können die Aussagen von ChatGPT direkt prüfen, selbst ausprobieren und so ihre Zufriedenheit mit den Antworten direkt bewerten.

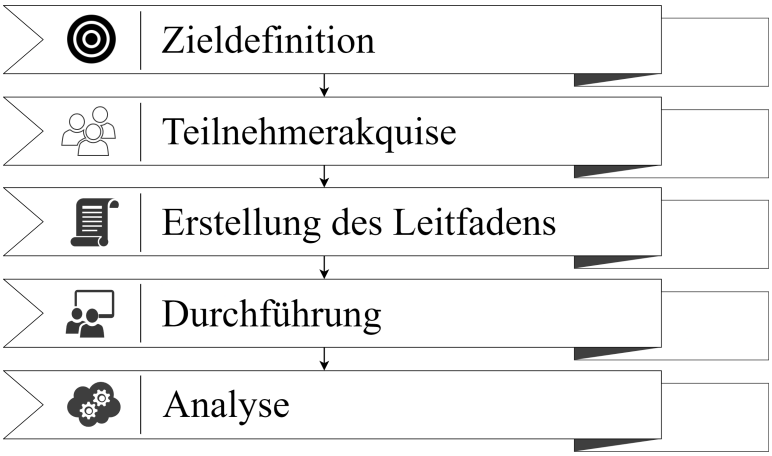


Abbildung 6.1: Ablauf der Interview-Studie

Der Teilnehmer konnte Fragen zu einem von ihm bestimmten Programm an ChatGPT stellen. Außerdem bekam der Nutzer eine vorgegebene Aufgabe, um diese mithilfe von ChatGPT zu lösen. Die Studie fand dabei in der gewohnten Arbeitsumgebung des Studienteilnehmers statt. Gewählt wurde die Form eines semi-strukturierten Interviews.

Zunächst wurden die Ziele der Interview-Studie herausgearbeitet und präzisiert, auf deren Grundlage der Interviewleitfaden erstellt wurde. Nachdem die Teilnehmer akquiriert wurden, erfolgte die Durchführung der Interviews. Abschließend erfolgte eine zusammenfassende Analyse (siehe Abbildung 6.1).

6.1.1 Zieldefiniton

Da bereits eine quantitative Studie durchgeführt wurde, war es wichtig, die genauen Ziele der nun folgenden qualitativen Studie festzuhalten. Dazu wurde das GQM Vorgehen angewendet. Folgende Ziel-Fragen und Unterfragen sollte die Interview-Studie beantworten:

- **G1**: Kann ChatGPT Erklärungen liefern, die der Nutzer nachvollziehen und durchführen kann bzw. ihm weiterhelfen?
  - **Q1**: Sind die Antworten korrekt?
  - **Q2**: Kann den Antworten gefolgt werden?
  - **Q3**: Wie präzise sind die Antworten? Wie sehr entsprechen sie dem, was der Nutzer meinte?
  - **Q4**: Welche Probleme ergeben sich bei der Nutzung?
- **G2**: Ist ChatGPT eine gute Alternative zum ‚Herumprobieren' bzw. Googeln?
  - **Q5**: Hat der Nutzer das Gefühl, schneller/leichter arbeiten zu können?
  - **Q6**: Ist es möglich, mit den Angaben von ChatGPT (neue) Aufgaben zu lösen?
- **G3**: Wie arbeitet es sich mit ChatGPT im realen Kontext?
  - **Q7**: Wie zufrieden sind die Nutzer?
  - **Q8**: Ist die Nutzung ablenkend (in irgendeiner Form)?
  - **Q9**: Ist die Nutzung zielführend?
  - **Q10**: Hilft ChatGPT beim Erlernen und Verstehen der Software?
- **G4**: Welche Erfahrungen sammelt der Nutzer durch und bei der Verwendung?

- **Q11**: Werden die Erwartungen des Nutzers erfüllt?
- **Q12**: Ändert sich die Nutzermeinung zu ChatGPT?
- **Q13**: Würde er die Nutzung weiterempfehlen?
- **Q14**: Würde er es als Softwareerweiterung nutzen wollen? Warum?
- **Q15**: Was ist die abschließende Meinung des Nutzers?

Die Fragen sind überwiegend subjektiv und werden daher als Frage im Interview aufgegriffen oder durch Beobachtungen festgehalten. Die Korrektheit und Präzision kann grob gezählt werden, indem die Aussagen von ChatGPT ausprobiert und so direkt geprüft werden. Umformulierungen der gestellten Frage geben Rückschlüsse darauf, ob die Intention des Nutzers getroffen wurde.

6.1.2 Teilnehmerakquise

Unter den Teilnehmern, welche an der Online-Studie teilgenommen hatten, wurde nach dem Interesse an der Teilnahme einer fortführenden Studie gefragt. Mit dem Ziel, fünf Teilnehmer für die Interview-Studie zu finden, wurden die Interessenten angefragt und zusammengestellt.

Teilnehmer

Die fünf interviewten Teilnehmer waren im Alter zwischen 20 und 35 sowie eine Person über 60. Vier der fünf Teilnehmer waren weiblich. Bei allen wurde die Studie am vertrauten, privaten Arbeitsplatz durchgeführt, in fast allen Fällen mit einem zur Verfügung stehenden zweiter Bildschirm für ChatGPT. Drei Nutzer gaben an, berufstätig zu sein, zwei Personen, sowohl zu studieren als auch berufstätig zu sein.

Fast alle Studienteilnehmer hatten wenig oder gar keine Erfahrung mit ChatGPT; der Teilnehmer mit der meisten Nutzung gab an, ChatGPT in letzter Zeit öfter verwendet zu haben und ansonsten geschätzt einmal im Monat. Das Vorwissen der Nutzer zu ChatGPT beschränkte sich in der Regel auf den Fakt, dass Fragen gestellt und mehr oder weniger präzise Antworten erhalten werden können. Zwei der Teilnehmer wussten, dass das Modell auf vielen Daten basiert trainiert ist.

Die Teilnehmer sehen ChatGPT insbesondere als Inspirationsquelle für erste Ideen, Formulierungen von Texten und um einen anderen Blickwinkel zu bekommen. Das Vertrauen ist eher gering und die Einstellung gegenüber ChatGPT kritisch; Fakten werden kontrolliert, Quellen geprüft und generell Aussagen hinterfragt. Es sei hilfreich, Fragen gut zu stellen und eigenes Hintergrundwissen mitzubringen. Eine Teilnehmerin hatte hohes Vertrauen in die Aussagen von ChatGPT, eine andere meinte, sie habe gar kein Vertrauen, glaube aber auch nicht, dass ChatGPT dumm sei.

Alle hatten die Erwartung, dass ChatGPT sie, teils sehr, teils potenziell, für diese Studie gut unterstützen kann und hilfreiche Antworten liefert, insbesondere bei der Bearbeitung der gestellten Aufgabe. Eine Teilnehmerin erhoffte sich, mehr über ChatGPT sowie auch über das erklärte Programm zu erfahren.

6.1.3 Erstellung des Leitfadens

Anhand der in der Zieldefinition erarbeiteten Fragen wurde der Interviewleitfaden erstellt. Das Interview sollte außerdem eine Einleitung zur Studie enthalten, Daten zum Teilnehmer aufnehmen und grundlegende Fragen zu seiner Erfahrung und Einstellung zu ChatGPT klären. Die Interaktion sollte möglichst nah an einem realistischen Anwendungsszenario sein. Um das zu erreichen, hatte der Nutzer zum einen die Möglichkeit, selbst mitgebrachte Fragen zu einer selbstgewählten Software an ChatGPT zu stellen. Zum anderen wurden zwei Aufgaben vorab erarbeitet, wovon der Nutzer eine wählen konnte. Mithilfe der vorgegebenen Aufgabe wurde sichergestellt, dass Interaktionen des Nutzers mit ChatGPT im Rahmen des Interviews beobachtet werden können.

Für die Interviews wurden ein Template für die Interviewerin angefertigt sowie eine Version, welche dem Teilnehmer zur Verfügung gestellt wurde (Abschnitt B.1). Die Idee des semi-strukturierten Interviews neben der Beantwortung der Fragen ist es, dem Teilnehmer die Möglichkeit zu geben, eigene Äußerungen zu treffen und spontan aufkommende Inhalte festzuhalten. Das Template für den Interviewer beinhaltet daher neben den eigentlichen Interviewinhalten noch Platz für die Antworten des Teilnehmers, Anweisungen für den Interviewer sowie Stellen zum Festhalten des Interviewkontexts und den Startzeiten für die einzelnen Abschnitte. Der Ablauf des Interviews orientierte sich dabei an Folgendem:

- *Einführung zur Studie:* Diesen Teil bekam der Teilnehmer bereits zur Anfrage zur Studienteilnahme. Hier wird das Ziel der Studie sowie der Ablauf beschrieben und angeregt, für die Studie eigene Fragen zu einer selbst gewählten Software festzuhalten. Außerdem wird ausdrücklich darauf hingewiesen, dass der Teilnehmer während der Studie alle Anfragen ausprobieren darf und er sich nicht ‚dumm' vorkommen muss.

- *Angaben zum Teilnehmer:* Zu Beginn der Studie wird der Kontext, in welchem die Studie stattfindet, festgehalten. Der Ort konnte vom Teilnehmer gewählt werden - ein gewohnter Arbeitsplatz oder ein gestellter Raum am Institut - und entsprechend der technische Aufbau. Wie in der ersten Nutzerstudie werden Daten zum Teilnehmer erfasst und erste Fragen über die Meinung zu ChatGPT gestellt. Hierbei wurde sich an den Fragen orientiert, welche in der Online-Studie gestellt wurden. Zunächst wurde die allgemeine Meinung bzw. Einstellung des

Teilnehmers zu ChatGPT festgehalten. Weiter folgten die Fragen zu dem Vertrauen des Nutzers in ChatGPT und wie oft er ChatGPT nutzen würde. Zusätzlich wurde erfragt, was der Teilnehmer über ChatGPT weiß und welche Erwartungen er an ChatGPT im Rahmen der Studie hat.

- *Eigene Fragen des Teilnehmers:* Sofern sich der Teilnehmer eigene Fragen zu einer Software seiner Wahl überlegt hatte, konnten diese an ChatGPT gestellt werden. Beobachtungen und Kommentare hinsichtlich **G1** wurden notiert, und abschließend die Qualitätsfragen zu **G2** und **G3** gestellt. Da dieser Teil optional war, wurde dieser zeitlich strikter begrenzt als die Anderen und in einem Rahmen von 10 bis maximal 20 Minuten abgehalten.

- *Gestellte Aufgabe:* Hier konnte der Teilnehmer eine vorgegebene Aufgabe zu der Software Excel oder Notion wählen, je nachdem, womit er sich mehr identifizieren konnte. In Notion sollte eine Rezeptsammlung erstellt werden. In Excel sollten zu vom Teilnehmer eingefügten Beispieldaten die Gesamtkosten aller vorkommenden Kategorien aufgestellt werden. Auch hier werden Beobachtungen und Kommentare hinsichtlich **G1** notiert, und abschließend die Qualitätsfragen zu **G2** und **G3** gestellt.

- *Abschlussfragen:* Zum Ende wurden verbliebene Qualitätsfragen zu **G4** gestellt und dem Teilnehmer die Möglichkeit gegeben, ein Resümee zu ziehen bzw. seine Meinung und Erfahrung zur Studie mitzuteilen.

Erarbeitung der gestellten Aufgabe

Bei der gestellten Aufgabe sollte sichergestellt werden, dass der Teilnehmer sich mit der Software und der Fragestellung identifizieren kann. Dem Teilnehmer wurde die Möglichkeit gegeben, zwischen einer ihm bekannten Software, welche er bereits genutzt hat, und einer potenziell für ihn unbekannten, neuen Software zu wählen. Angepasst an die Probanden wurden Excel, als bekanntes Programm für Tabellenkalkulationen, und Notion, als den Teilnehmern unbekanntest Programm zur Organisation von Wissen und Projekten, gewählt. Zur konkreten Aufgabenfindung konnte sich dabei wieder an dem erfassten Erklärungsbedarf orientiert werden, da dem Teilnehmer hier die Wahl überlassen wird.

Für Excel sollte der Teilnehmer im Optimalfall mit Pivot-Tabellen arbeiten, ohne dass dies bereits in der Aufgabenstellung vorweggenommen wird. Die Aufgabe lautete daher:

> *Erstelle eine Excel-Tabelle, in der Name, Kategorie sowie Kosten von ausgedachten Anschaffungen beinhaltet sind (min. 50 Einträge). Lasse dir die Gesamtkosten für alle Kategorien berechnen.*

Die Aufgabenstellung ist bewusst vage und ohne nähere Erläuterung, um die Problemstellung realistisch zu halten. Hinzu kamen optionale Zusatzaufgaben zur weiteren Datenverarbeitung, welche darunter aufgelistet wurden.

Die Menge der Einträge sollte den Teilnehmer dazu veranlassen, ChatGPT nach Beispieldaten zu fragen. Wenn der Nutzer nicht auf die Idee kam, kann der Interviewer eingreifen und einen Hinweis darauf geben. Dies geschieht jedoch erst, nachdem der Teilnehmer dabei ist, erste Einträge zu erstellen. Der Hinweis soll zum einen verhindern, dass unnötig Zeit an der repetitiven Arbeit verbracht wird und zum anderen den Teilnehmer einen ersten Anstoß geben, wie ChatGPT verwendet werden kann. Dabei soll die eigentliche Lösung der Aufgabe in Excel stattfinden und ChatGPT nur als Hilfe bzw. Unterstützung zur Lösung des Problems mit Excel genutzt werden. Der Teilnehmer soll sich vorstellen, er hätte einen vorgegebenen Datensatz, auf welchem er Berechnungen anstellen soll. Erlaubt ist also, von Excel Daten in ChatGPT zu kopieren, nicht jedoch direkt basierend auf den zuvor erstellten Einträgen direkt die Gesamtkosten von ChatGPT berechnen zu lassen.

Für Notion wurde eine kreativ freie Aufgabe gestellt:

> *Erstelle in Notion (die Struktur für) eine oder mehrere Rezeptsammlungen mit Back- und Kochrezepten nach deinen Vorstellungen*

Der Teilnehmer musste sich hier also entweder selbst eine Strukturherangehensweise ausdenken, oder, wie angenommen, ChatGPT nach einem geeigneten Vorgehen fragen. Die Unteraufgaben bauten darauf auf, dass der Nutzer eine Datenbank erstellt sowie eine entsprechende Ansicht auf der Seite anzeigen lässt. Zuvor waren verschiedene Ideen gesammelt worden, was der Teilnehmer erstellen sollte. Möglichkeiten waren die Erstellung einer Überblicksansicht mit der Organisation von verschiedenen Unterseiten wie To-do Listen, ein Tagebuch, Einkaufslisten, Monatsplanung und Rezeptsammlungen. Da jedoch bereits das Erstellen einer dieser Punkte für Notion-Unerfahrene eine Herausforderung war (siehe Testlauf), wurde sich auf die Rezeptsammlungs-Seite mit möglichen Unterseiten beschränkt.

Testlauf und ChatGPT Anpassung

Zu beiden Aufgaben wurde ein Testlauf durchgeführt, um die ungefähre Dauer zur Lösung des Aufgabenauftrags zu bestimmen und die Machbarkeit beispielhaft zu prüfen. Außerdem wurden verschiedene Einstellungen der

ChatGPT-Antwort-Präferenzen getestet. Unter der Einstellung *„Customize ChatGPT“* konnte angegeben werden, in welcher Weise ChatGPT antworten solle, welche dann für jeden neuen Chat angewandt wurden. Das Upgrade zu ChatGTP-4 und eigenen Modellen stand zu diesem Zeitpunkt nicht mehr zur Verfügung, weshalb dieser Ansatz gewählt wurde. Diese Einstellung, oder die vorzuziehende Verwendung eines selbst erstellten Modells, ist realitätsnah, da bei einer Anbindung an eine Software ein spezialisiertes Modell verwendet werden sollte. Zunächst war die Angabe nur *„short and accurate“* zu antworten, was zu sehr verkürzten Antworten führte. Auch noch mit dem Hinweis, falls nötig ausführlicher antworten zu dürfen. Folglich wurde danach gebeten, Rückfragen zu stellen, wenn ChatGPT genauere Informationen benötigte. Wird beispielsweise gefragt, wie in Notion eine Liste angelegt wird, sollte ChatGPT zurückfragen, welche Art von Liste der Nutzer erstellen möchte. Obwohl dies zunächst funktionierte und in der finalen Liste der Anpassungen enthalten war, wurden Rückfragen nur bedingt bis gar nicht gestellt. Es wurde getestet, eine Antwort nach der nächsten ausgeben zu lassen, da die Anleitungen recht lang ausfielen. Dies führte jedoch dazu, dass kein Gesamtüberblick über die Anleitung mehr möglich war bzw. nicht direkt erkenntlich war, ob die Schritte auf das gewünschte Ziel hinauslaufen. Daher wurde dieser Ansatz wieder verworfen. Stattdessen sollte ChatGPT die einzelnen Schritte mit einer Beschreibung, was und wie zu tun ist, auflisten. Zum Schluss wurde das *„short and accurate“* umgewandelt in „precise but short“, um potenziell längere Antworten erhalten zu können.

Letztendlich wurden folgende Konfigurationen für die Antworten von ChatGPT verwendet:

> *„answers should be precise but short. If it is something you generally can say more about, its ok for the answer to be longer. ChatGPT can ask questions back to get more into detail. If the question is asking for a way to do something the answer should be a guide to do so. The description of the step should contain not only what to do but also how to do it“*

Beim Herausarbeiten und Testen der Excel-Aufgabe im Vorhinein wurden etwa 40 Minuten aufgewandt, ohne alle optionalen Schritte durchzuführen. Zunächst war hier auch nach Filtern und Sortieren gefragt - aufgrund des Zeitumfangs wurde dies aus der Aufgabenstellung entfernt. Bei der Frage *„Wie kann ich mir die Gesamtkosten für die jeweiligen Kategorien anzeigen lassen?“* schlug ChatGPT wie gewünscht als Alternativ-Weg zur Summenformel Pivot-Tabellen vor.

Eine ähnliche Zeit wurde für die Notion Aufgabe verwendet. Dabei galten die gleichen Voraussetzungen wie für den Teilnehmer; die Software

wurde zuvor noch nicht genutzt. Problematisch war hierbei, dass die Oberfläche von Notion auf Englisch war und die Begriffe in ChatGPT auf Deutsch. Bei dem Hinweis, dass in Notion die Begriffe alle auf Englisch sind, antwortete ChatGPT komplett auf Englisch, anstelle nur die einzelnen englischen Begriffe zu verwendet. Wie sich nach den Interviews herausstellte, wäre eine Umstellung der Oberfläche möglich gewesen. Auf diese Option gab es keinen Hinweis durch ChatGPT. Auch ohne die Sprachunterschiede waren die Begriffe nicht immer konsistent verwendet oder Vorgehen ungenau erklärt. Das mentale Modell zur Webanwendung und ihren Funktionen unterschied sich von den tatsächlichen Ver- und Anwendungsmöglichkeiten der Software. Beispielsweise wurde für die Erstellung eines Rezeptbuchcovers versucht, ein Text über ein Bild zu ziehen. Das scheint in Notion jedoch, entgegen ChatGPTs Vorschläge, nicht möglich zu sein.

6.1.4 Durchführung

Die Interviews wurden Ende März 2024 innerhalb einer Woche mit der kostenfreien Version von ChatGPT-3.5 durchgeführt. Da alle Teilnehmer sich für das Homeoffice entschieden hatten, wurde die Studie mit dem PC oder Laptop des Teilnehmers ausgeführt. Eine Überlegung war, ChatGPT auf dem Laptop der Interviewerin zur Verfügung zu stellen. Diese Idee wurde verworfen, da so der Wechsel zwischen der Verwendung der Software und ChatGPT noch größer und entfernter von einer direkten Softwareanbindung war. So sollte zu Beginn der Studie der Teilnehmer ChatGPT und Excel bzw. Notion je nach Aufgabenwahl öffnen, damit die Interviewerin sich mit dem Universitäts-Account einloggen konnte.

Der Teilnehmer konnte sich seine Arbeitsumgebung selbst einrichten. Wenn möglich zogen die Teilnehmer ChatGPT auf den zweiten Bildschirm, um beide Programme gleichzeitig gut benutzen zu können. Währenddessen hielt die Interviewerin den Kontext bzw. Aufbau fest, mit wem und wann die Studie durchgeführt wurde und legte den Teilnehmerbogen vor.

Dann startete das eigentliche Interview. Angaben wurden stichpunktartig festgehalten. Innerhalb der ersten 5 bis 10 Minuten wurden Informationen zum Teilnehmer festgehalten und über seine Erfahrungen und Erwartungen gesprochen. Zwei Teilnehmer hatten eigene Fragen vorbereitet, von welchen die ersten ein bis zwei gestellt wurden. Hier wurde bei einem Teilnehmer der maximale Zeitrahmen von 20 Minuten benötigt, um eine mehr oder weniger ihn zufriedenstellende Antwort zu erhalten.

Es folgte die Aufgabe, für welche 20 bis 40 Minuten vorgesehen war. Der Teilnehmer konnte nach eigenem Ermessen entscheiden, ob er diesen Zeitrahmen überschreiten und wie weit er die optionalen Aufgaben noch bearbeiten wollte.

Eine Teilnehmerin entschied sich für Notion, da sie etwas Neues kennenlernen wollte. Aufgrund eines an dieser Stelle nicht testkonformen Ansatzes war es nicht möglich, die Aufgaben weiter zu bearbeiten; deshalb wurde diese Aufgabe nach knapp 40 Minuten unterbrochen. Die Teilnehmerin erstellte keine Datenbank sondern einzelne Unterseiten, welche sie gerne im Nachhinein zusammengefasst hätte.

Die anderen vier Teilnehmer entschieden sich für Excel, mit der Begründung, dass ihnen Notion nichts sagt bzw. mehr Interesse daran bestand, die ihnen bekannte Software Excel zu verwenden. Jeder Teilnehmer gab an, mit Excel bereits gearbeitet zu haben. Außerdem hatte jeder der Vier die Software installiert, sodass auf diese Version zurückgegriffen werden konnte. Die Teilnehmer sollten die Aufgaben selbständig bearbeiten, auch wenn teilweise die Interviewerin um Rat gefragt wurde. Mit dem Hinweis, dass ChatGPT gefragt werden könne, hielt sich die Interviewerin weitestgehend zurück. Nur bei größeren Problemen oder, um bei der Erstellung der anfänglichen Beispieleinträge Zeit zu sparen, griff die Interviewerin ein. Die Voraussetzung dazu war, dass der Teilnehmer ausreichend Zeit hatte, eigene Überlegungen auszuprobieren. Schon kleine Tipps zum Umgang mit ChatGPT halfen den Teilnehmern sehr weiter. Im Gegensatz zum Testlauf schlug ChatGPT beispielweise bei keinem der Teilnehmer direkt die Verwendung von Pivot-Tabellen vor, sodass hier erst nach längerem Herumprobieren ein Tipp von der Interviewerin kam, nach alternativen Möglichkeiten zu fragen.

Im Anschluss an die Bearbeitung der eigenen Fragen sowie der Aufgabe wurden die Teilnehmer zu ihrer Wahrnehmung und Zufriedenheit der Antworten von ChatGPT befragt. Abschließend waren weitere 10 Minuten vorgesehen, um über die nun gesammelten Erfahrungen des Teilnehmers zu sprechen und inwieweit er ChatGPT weiterempfehlen oder nutzen wollen würde. Außerdem konnte er eine abschließende Meinung äußern und es wurde ihm für seine Teilnahme gedankt.

6.1.5 Analyse

Die Studie wurde manuell ausgewertet. Dabei wurden über mehrere Teilnehmer hinweg öfter vorkommende oder genannte Punkte als besonders wichtig eingestuft. Auf dieser Basis wurden Hauptergebnisse zusammengestellt. Ebenso wurde das beobachtete Nutzerverhalten analysiert und in die Auswertung mit aufgenommen.

6.2 Ergebnisse Interview-Studie

Die Zufriedenheit der Nutzer mit den Antworten von ChatGPT war zwiegespalten. Teilweise wurde es als schneller angesehen mit den Antworten von ChatGPT zu arbeiten als andere Quellen zu nutzen. Manche Nutzer merkten an, dass sie sich vorstellen könnten, dass bei bestimmten Problemen ein Video oder eine bildliche Anleitung hilfreicher gewesen wäre. Die Arbeit und das Lösen der Aufgabe wurde durch die Erklärungen von ChatGPT bedingt erleichtert. Je nach Stellung der Frage und des Vorwissens des Nutzers bezüglich der Software und des Problemhintergrundes sei teilweise Arbeit abgenommen worden. Nutzer resümierten, dass ChatGPT „wenig hilfreich sei, wenn man gar keine Ahnung von der Software hat“.

Von der Interviewerin gegebene Hinweise bzw. erste Tipps zur Nutzung von ChatGPT seien sehr hilfreich gewesen, um mit angepassten Prompts bessere Antworten zu erhalten. Eine Beobachtung der Interviewerin war, dass die Teilnehmer dennoch teilweise, wenn sie nicht weiter wussten, selbst über Unklarheiten rätselten und die Interviewerin anstelle von ChatGPT nach Rat fragten.

Umfangreiche Erklärungen trugen zu der Verwirrung der Teilnehmer bei. Letztlich merkten mehrere Teilnehmer an, dass sie sich aufgrund des nicht direkten Verstehens der Erklärungen „dumm vorkamen“. Weiter fiel es Teilnehmern schwer, von ChatGPT vermitteltes Wissen zu erfassen und aktiv anzuwenden. Ursächlich dafür waren unter anderem auch von ChatGPT teilweise falsch oder unvollständig generierte Vorschläge zur Lösung der Aufgabe. Im Beispiel von Excel konnte z.B. der von ChatGPT generierte Code nicht direkt ohne Anpassung vom Teilnehmer in seine Arbeit übernommen werden. Die Formel war teils allgemein gehalten und nicht auf das Anwendungsbeispiel, welches der Nutzer gegeben hatte, zugeschnitten.

Fast alle Nutzer tendierten dazu, bei Klickanleitungen die ersten Schritte bzw. Zeilen zu lesen und dann aufzuhören. Wichtige erklärte Details werden so überlesen. Die Erklärungen wurden als „unnötig“ lang angesehen und wären hilfreicher, wenn die deutlich kürzer und präziser ausfallen würden. Bei Nachfragen vom Nutzer zu einzelnen Punkten verlor dieser schnell den Überblick im Chatverlauf über abgearbeitete, geänderte und noch durchzuführende Schritte.

Über die Interviews hinweg konnte die Interviewerin beobachten, dass sich die Antwortengebung von ChatGPT sehr unterschied. Einer Teilnehmerin wurde zur Berechnung der linearen Regression eine nach Excel kopierbare Tabelle generiert. Die enthielt die wichtigsten Faktoren zur Abschreibung

sowie Formeln in den einzelnen Zellen zur Berechnung der einzelnen Werte. Die Abhängigkeiten der Zellen untereinander wurden dabei nicht korrekt generiert. Anderen Teilnehmern wurde der Vorgang unterschiedlich abstrakt erklärt und eine Abschreibungstabelle hätte vollständig vom Teilnehmer selbst angelegt werden müssen. Im Gegensatz zum Testlauf schlug ChatGPT keinem der Teilnehmer von sich aus die Verwendung von Pivot-Tabellen bei dem beschriebenen Problem vor.

Die Nutzer gaben einstimmig an, dass die Benutzung von ChatGPT selbst sie nicht abgelenkt habe. Manche Nutzer meinten, sie hätten durch die Verwendung von ChatGPT etwas Neues über die ausgewählte Software gelernt. Insgesamt wurde ChatGPT von den Teilnehmern als hilfreich bewertet, insbesondere für das Erhalten eines ersten Ansatzes. Obwohl die Erwartungen der Teilnehmer von ChatGPT großteils nicht erfüllt wurden, konnten sie sich vorstellen, ChatGPT als Software Integration unterstützend zu nutzen. Die Teilnehmer würden ChatGPT einschränkt weiterempfehlen, unter der Bedingung, dass mit den Informationen kritisch umgegangen und Quellen geprüft werden.

Abschließende Meinungen waren, dass die Studie und die Verwendung von ChatGPT Spaß gemacht hat. Das Arbeiten im Rahmen der Studie sei interessant gewesen. Neue Erkenntnisse bezüglich des Umgangs mit und Verwendung von ChatGPT aber auch stellenweise mit der Software wurden erlangt.

Kapitel 7

Sonstige Auswertungen zu Metriken

7.1 Lesbarkeitsindizes

Zur Bewertung der Verständlichkeit wurden die Lesbarkeitsindizes des Flesch Reading Ease Scores und KWMs ausgewertet, bzw. im Falle des KWMs die Variablen mit dem höchsten Einfluss auf den KWM (Abschnitt 3.2).

Flesch Reading Ease Score

Der Lesbarkeitsindex des Flesch Reading Ease Scores wurde mittels der Bibliothek textstat (Version 0.7.3) in Jupyter Notebook ausgewertet. Die Bibliothek nutzt die in den Grundlagen angegebene Formel für die deutsche Sprache (Abschnitt 2.3). Hier wurde für alle Erklärungen *Antworten 1* und *Antworten 2* von ChatGPT der Lesbarkeitsindex ermittelt und Vergleiche gezogen. Ergebnisse, welche unter dem Wert 0 lagen wurden entsprechend des Wertebereichs auf 0 gesetzt. Die Werte für die *Antworten 1* lagen zwischen 0 und 72,05 (M = 38; SD = 17,27), für die *Antworten 2* zwischen 0 und 78,25 (M = 37,8; SD = 17,39). Null entspricht dabei einer sehr schlechten Lesbarkeit, die Werte ab 70 einer mittelleichten, aber gut verständlichen Lesbarkeit.

Antworten 1 wiesen zu den folgenden Prompts eine Lesbarkeit von 0 (oder niedriger berechneter) auf: 75 (x), 76, 120, 183 und 217.

Die mit (x) markierten Prompts und Antworten kamen in der durchgeführten Online-Studie vor. Auch bei den Antworten 2 erhielt der Prompt 75 eine Lesbarkeit von 0, ebenso wie die folgenden weiteren: 75 (x), 95, 98, 121, 135 (x), 176 (x), 184 (x) und 212.

Prompt 75 erhielt in der Online-Studie eine gemittelte Wertung der Teilnehmer zwischen *neutral* und *etwas* verständlich (M = 3,58; SD = 0,97; *Antwort 1*: M = 3,42; SD = 0,90; *Antwort 2*: M = 3,75; SD = 1,06).

- Promt 75: *„Wie kann ich bei Eclipse die Hibernate-Anbindung richtig konfigurieren?“*
  - Antwort 1: *„Um Hibernate in Eclipse zu konfigurieren, musst du das Hibernate-Tool-Plugin installieren, eine Hibernate-Konfigurationsdatei (hibernate.cfg.xml) erstellen und deine Entity-Klassen definieren.“*
  - Antwort 2: *„Füge die Hibernate-JARs zu deinem Projekt hinzu und erstelle eine Hibernate-Konfigurationsdatei (hibernate.cfg.xml). Konfiguriere deine Datenbankverbindung und Entity-Klassen in dieser Datei.“*

Die Prompts mit der besten berechneten Lesbarkeit waren Prompt 63 (72,05) und Promt 68 (71,9) für die jeweilige *Antwort 1*. Der Prompt 63 erhielt in der Online-Studie eine mittlere Bewertung der Verständlichkeit von M1 = 4,46 (SD = 0,78) und M2 = 4,35 (SD = 1,06), wobei 5 der höchsten Zustimmungstufe *sehr* verständlich entspricht.

- Prompt 63: *„Wie erstelle ich eine eigene zoom Veranstaltung?“*
  - Antwort 1: *„1. Melde dich bei deinem Zoom-Konto an. 2. Klicke auf „Meetings“ und dann auf „Meeting planen“. 3. Gib die Details deines Meetings ein (Titel, Datum, Uhrzeit, etc.). 4. Wähle ggf. zusätzliche Optionen wie Passwort oder Wartezimmer. 5. Klicke auf „Speichern“, um das Meeting zu planen. Teile den Meeting-Link mit den Teilnehmern.“*

Respektive waren die höchsten Lesbarkeitsindizes bei den *Antworten 2* der Prompt 131 (78,25) und der Prompt 14 (74,4).

- Prompt 14: *„Wie kann man in Excel in mehrere Zellen gleichzeitig dasselbe eintragen?“*
  - Antwort 2: *„Markiere die gewünschten Zellen, gib den Eintrag ein und drücke ”Strg” + ”Enter”. Dies trägt den Text oder die Formel in alle ausgewählten Zellen ein.“*

Prompt 14 kam auch in der Online-Studie vor und erhielt da mittlere Bewertungen der Verständlichkeit von M1 = 4,5 (SD = 0,81) und M2 = 4,88 (SD = 0,34), wobei 5 der höchsten Zustimmungstufe *sehr* verständlich entspricht.

KWM

Die betrachteten Variablen des KWM (Abschnitt 2.3) waren die Verbintensität sowie die Satzlänge.

Für die Bestimmung der Verbanzahl wurde zunächst versucht, diese mittels der nltk Bibliothek in Jupyter Notebook ermitteln zu lassen, mit deren Hilfe auch die Satzlänge bestimmt wurde. Die Genauigkeit zur Erkennung der Verben war jedoch gering und viele Verben wurden nicht als solches gezählt. Es wurde daher entschieden, die Verben für die in der Studie vorkommenden Erklärungen zu den Prompts manuell auszuzählen. Dabei wurde sich nach dem Vorgehen von Briest [3] zur Bewertung der Verben gerichtet. Gezählt wurde die Anzahl aktiver Verben. Entsprechend wurden Zusammensetzungen eines Verbs mit haben, werden, würden als ein Verb gezählt. Kam ein Verb im Satz mit der Verbindung eines Verbs wie wollen, sollen oder können vor, so wurden diese als zwei Verben getrennt gezählt.

Ein Beispiel wird anhand der zu dem Prompt 143 *„Wie kann ich bei Tiktok Videos wiederfinden?“* gegebenen *Antwort 1* verschaubildlicht:

> *„Wenn du ein Video bei TikTok wiederfinden möchtest, kannst du die Suchfunktion nutzen und Stichworte eingeben. Wenn du ein Video geliked hast, findest du es auch unter deinem „Gefällt mir“-Tab in deinem Profil.“*

Wiederfinden möchtest wird entsprechend als zwei Verben gezählt, und *geliked hast* als ein Verb. Damit ergibt sich für die Beispielerklärung eine Verbanzahl von sieben sowie eine Verbintensität von sieben-durch-zwei-Sätzen, also 3,5.

Es ergaben sich die in Tabelle 7.1 dargestellten Mittelwerte. Zu erkennen ist, dass sich die Antworten in der Summe in den einzelnen Parametern kaum unterscheiden.

Mittelwert der	**Antworten 1**	**Antworten 2**
Satzlänge (# Wörter)	15,62	15,16
# Sätze	2,14	1,96
# Verben*	4,68	4,41
# Verben/Satz*	2,35	2,4

Tabelle 7.1: Mittelwerte Sätze und Verben der beiden generierten Erklärungen (*nur für Prompts aus der Online-Studie)

7.2 Manuelle Auswertung verbliebener Metriken

Bei der Durchführung der Interview-Studie wurden Inkonsistenzen und Inkorrektheiten in den Antworten von ChatGPT erfasst. Durch eine manuelle Auswertung der Daten wurden ergänzende Betrachtungen gesammelt.

Probleme bei der Präzision traten insbesondere bei Schrittanleitungen auf. Dort wurden einzelne Schlagworte unterschiedlich stark abweichend benannt. Ebenso konnten einzelne Schritte so oder in so ähnlicher Form nicht wie beschrieben, sondern unter einem anderen Ort wiedergefunden werden.

Beispielsweise wurde für das Einfügen von Excel-Daten in Word folgende Anleitung ausgegeben:

> *„[...] du kannst Daten aus Excel in Word einfügen, indem du sie in Excel kopierst und in Word mit der Option „Einfügen“ > „Inhalte einfügen“ > „Verknüpfen und Daten einfügen“ einfügst.“*

Hier ist zunächst unklar, welcher Optionsaufruf genau gemeint ist. Zum einen gibt es den Reiter „Einfügen“, in welchem „Inhalte einfügen“ oder eine Verknüpfung nicht zur Auswahl steht. Mit Rechtsklick erscheinen mehrere Möglichkeiten zum „Einfügen“. Der folgende Schritt heißt in Word nicht „Inhalte einfügen“, sondern „Einfügeoptionen“. In diesem Reiter kann aus unterschiedlichen Unterpunkten eine Verknüpfungsoption gewählt werden. „Verknüpfen und Daten einfügen“ ist nicht aufgelistet.

Die Verwendung von Füllwörtern und Abkürzungen wurde ebenfalls manuell überprüft. Abkürzungen wurden bis auf „z.B.“ und „etc.“ nicht verwendet. Diese stellen allgemein gebräuchliche Abkürzungen dar.

Bei der Verwendung von Füllwörtern fiel auf, dass ChatGPT insgesamt sehr wenig verwendet. Ausnahmen sind, wenn diese zur Betonung oder Verstärkung verwendet werden, wie „jedoch, auch, allerdings, aber“. Bei Verallgemeinerungen, „Vermutungen“ bzw.„Unsicherheiten“ seitens ChatGPT treten im Vergleich vermehrt Füllwörter auf.

Kapitel 8

Diskussion

In diesem Kapitel werden die in Unterabschnitt 3.1.2 aufgestellten Forschungsfragen anhand der zusammengetragenen Ergebnisse dieser Arbeit untersucht und beantwortet. Außerdem werden die aufgestellten Hypothesen geprüft und es wird auf mögliche Limitationen der Arbeit eingegangen.

8.1 Beantwortung der Forschungsfragen

Ziel der Arbeit war es, herauszufinden, inwieweit sich ChatGPT zur automatisierten Generierung von Erklärungen als Einbindung für Software Systeme eignet. Die dazu aufgestellte Forschungsfrage lautete:

> ***RQ1:*** *Inwiefern kann eine generative KI wie ChatGPT prägnante Erklärungen zu verschiedenem Erklärungsbedarf und unterschiedlichen Software Systemen erzeugen?*

ChatGPT ist in der Lage, zu den hier untersuchten Software Systemen in vielerlei Hinsicht gute Erklärungen zu liefern. Folgendes gilt für die Online-Studie: Die überwiegende Mehrheit der von ChatGPT erzeugten Antworten ist in Hinblick auf Effizienz, Angemessenheit, Zufriedenheit und Effektivität zufriedenstellend bewertet. Die Verständlichkeit wurde im Vergleich dazu noch höher bewertet. Der Lesbarkeitsindex bestätigte weitestgehend die Bewertungsgüte der Antworten. Im Hinblick auf den Erklärungsbedarf, welcher in die Arten Domainwissen, Integration, Security und Systemverhalten unterschieden wurde, gab es keine signifikanten Unterschiede. Einzig bei der Erklärungsart der Warum-Erklärungen wurden durchschnittlich im Vergleich schlechtere Ergebnisse erzielt. Dies lag vor allem an der eingeschränkten Antwortlänge.

In der Interview-Studie gab es folgende markante Unterschiede: Nachvollziehbarkeit bzw. Verständlichkeit der Erklärungen wurde dort teilweise bemängelt. Die Erklärungen seien insbesondere weniger hilfreich und gut zu verstehen, wenn beim Nutzer nicht ausreichend Grundkenntnisse zur Software bzw. Fachkenntnisse vorliegen. Korrektheit wurde in beiden Studien angezweifelt, in der Interview-Studie wurde festgestellt, dass Erklärungen zu unpräzise oder falsch waren. Trotz dieser Mängel wurden die Nutzer durch die Hinweise von ChatGPT in die richtige Richtung gelenkt. Insgesamt konnten sich die Nutzer vorstellen, dass eine Einbindung auf Dauer zu besseren Antworten und einfacherer Systemnutzung führen wird. Auch die Erfahrung, mit ChatGPT zu arbeiten, würde die Effizienz und Effektivität steigern.

Interessant sind die unterschiedlichen Aussagen beider Studien. Diese könnten zustande gekommen sein, weil ChatGPTs Antworten oberflächlich gesehen gut erscheinen, in der Tiefe jedoch Mängel aufweisen. Dies wurde auch durch die manuellen Untersuchungen bestätigt.

> ***RQ2:*** *Inwiefern beeinflussen die Erklärungen die Nutzung bzw. Bedienung und das Verständnis des Nutzers gegenüber der Software?*

Da in der Online-Studie der Einfluss nur erahnt und geraten werden konnte, wurde eine weitere Studie zur Klärung der Forschungsfrage RQ2 durchgeführt. Die Teilnehmer der Interview-Studie gaben an, dass die Erklärungen von ChatGPT das Arbeiten potenziell schneller und einfacher gestalten können. Allerdings sei dies sehr abhängig von der eigenen Präzision beim Formulieren der Prompts sowie dem eigenen Vorwissen zur Software. Es ist deutlich einfacher, ChatGPT gezielt in eine Richtung zu befragen, als wenn kein oder wenig Grundwissen zur Lösung der Aufgabe besteht. Die gesammelten Erfahrungen des Autors ergänzen diese Ansicht.

> **RQ3:** Welche Aspekte der Erklärbarkeit sind besonders wichtig für den Nutzer?

Für die Nutzer spielte in erste Linie eine gute Verständlichkeit und damit verbundene Einfachheit der Antworten eine Rolle. Eine klare Struktur bzw. Aufbau der Erklärungen sowie eine angemessene Ausführung in Länge und Tiefe sind den Nutzern besonders wichtig.

Wünschenswert ist ebenso eine ergänzende Erklärungsmöglichkeit anhand eines konkreten, zur Nutzerfrage passenden Beispiels. In der Antwort gegebene Excel-Formeln konnten nicht eins zu eins übernommen werden, sondern musste oftmals noch entsprechend angepasst werden. Bei einigen Antworten mangelte es an Korrektheit. Die Interviewerin beobachtete eine daraus stark resultierende, negative Beeinflussung bei der Arbeit der Nutzer. Der Nutzer wurde dennoch in den meisten Fällen in die richtige Richtung gelenkt und konnte sich durch fehlerhafte Versuche zur Lösung durcharbeiten.

Abschließend wurde es insgesamt als wichtig erachtet, dass ChatGPT als erster Hinweisgeber gut verwendbar ist und zu ersten Ideen zur Lösung verhilft.

RQ4: Welche weiteren Faktoren haben Einfluss auf die Güte der Erklärungen?

Die Ausführlichkeit der Antworten hatte einen großen Einfluss auf die Bewertung der Erklärung durch den Nutzer. Unangemessen zu lange Antworten bürgen die Gefahr, nicht gründlich gelesen zu werden und den Überblick zu verlieren. Bei Anfragen die durch Zwischenfragen notgedrungen geschachtelt wurden, ist dieser Effekt besonders auffallend. Die zuvor generierten Antworten der höheren Level werden meist nicht mehr berücksichtigt. Die Aufmerksamkeitsspanne des Nutzers wird zu stark beansprucht und die generelle Lesebereitschaft sinkt merklich.

Erheblichen Einfluss haben neben der Vorerfahrung des Nutzers mit der Software und seinem zur Problemstellung gehörende Fachwissen vor allem der sichere Umgang mit ChatGPT und die Formulierung der Prompts.

Die „Zufälligkeit“ im Sinne von nicht identischen Formulierungen, unterschiedlicher Wortwahl, Satzaufbau und Strukturwahl bei gleichen Anfragen führt zu einer Varianz von Antworten; wobei die Antworten mal besser mal schlechter bewertet wurden.

Je nach dem, was für eine Art Frage gestellt wurde, wird die ChatGPT Antwortgüte von den Nutzern unterschiedlich bewertet. Die zufriedenstellende Beantwortbarkeit der Frage ergibt sich häufig auch aus dem Fragetypus selbst und den dadurch höheren Schwierigkeitsgrad.

Alle Teilnehmer der Interview-Studie neigten dazu, an sich selbst zu zweifeln, sich „dumm zu fühlen“ und somit eher bei sich die Schuld zu suchen, anstatt die Güte der Erklärung unvoreingenommen zu beurteilen. ChatGPT gelang es offenbar nicht, für zugängliche Nachvollziehbarkeit und Verständnis zu sorgen. Die Nutzer hatten ab und zu den Eindruck, die Erklärung nicht verstanden zu haben und nicht in eigenen Worten wiedergeben zu können. Die Gefahr den Anweisungen blind zu folgen wurde durchaus gesehen.

8.2 Betrachtung der Hypothesen

H1: ChatGPT ist in der Lage, gute Erklärungen für verschiedenen Erklärungsbedarf und unterschiedliche Art von Nutzern zu generieren *(RQ1, RQ4)*
Wie sich in den Nutzerstudien zeigte, ist ChatGPT generell in der Lage, auf viele Fragen gute Erklärungen für Nutzer unterschiedlichen Vorwissens, Alters, Geschlechts und Berufshintergrunds sowie verschiedenem Erklärungsbedarf zu generieren. Hierbei gibt es jedoch auch kontroverse Meinungen, welche die Erklärungen als nicht ausreichend oder nicht ganz richtig ansahen. Wie gut die Erklärungen empfunden werden, ist abhängig von mehreren Faktoren, welche bei der Klärung der Forschungsfragen festgehalten wurden.

H2: Die Aussagen von ChatGPT können als automatisierter Ersatz zur manuellen Entwicklung von statischen Erklärungen verwendet werden *(RQ1)*
Die Erklärungen von ChatGPT sind in der Lage, verschiedene Fragen zu Software Systemen zufriedenstellend zu erklären. Auch wenn nicht jede Frage des Nutzers nach seinen Vorstellungen beantwortet wird, so bietet ChatGPT den Vorteil, Erklärungen an den Bedarf angepasst zu generieren. Entwicklern würden die Ressourcen zur Erhebung und Entwicklung von Erklärungen abgenommen.

H3: ChatGPTs Antworten sind überwiegend korrekt oder mindestens richtungsweisend *(RQ1)*
In der Korrektheit zeigen sich Mängel, nicht nur in ihrer eigentlichen Richtigkeit, sondern auch in ihrer Vollständigkeit und Ausführlichkeit. So werden Fragen nur teils (richtig) beantwortet, geben dadurch Nutzern jedoch einen richtungsweisenden Ansatz.

H4: Dem Nutzer wird durch die Antworten von ChatGPT die Bedienung der Software erleichtert *(RQ2)*
Die Ergebnisse der Online-Studie weisen zwar ein positives Feedback auf, jedoch erwies sich in der Interview-Studie die Anwendbarkeit der Erklärungen stellenweise als schwierig. Die Interview-Studie ergab, dass ChatGPT wenig hilfreich ist, wenn dem Nutzer die Software gänzlich unbekannt ist. Je mehr Vowissen der Nutzer zu der Software als auch zu ChatGPT mitbrachte, desto hilfreicher waren, auch unpräzisere, Erklärungen. Als Hilfestellung bei komplizierteren Aufgaben sei ChatGPT nicht besonders hilfreich.

H5: Der Arbeitsfluss wird nicht unterbrochen, sondern unterstützt *(RQ2)*
Hier war das Feedback sehr eindeutig. Nutzer sahen keinen negativen Einfluss von ChatGTP auf ihren Arbeitsfluss, obwohl ChatGPT in der Interview-Studie in einem separaten Fenster bedient werden musste.

H6: ChatGPT stellt Spekulationen an, die als solches auf Anhieb nicht erkennbar sind *(RQ1, RQ2, RQ4)*
Vor allem in der Interview-Studie zeigte sich, dass ChatGPT bei dem Betiteln von Funktionen wie Formeln in Excel dazu neigte, sich Namen 'auszudenken'. Erst durch das direkte Ausprobieren zeigte sich, wenn Angaben nicht richtig waren. Auch in der Online-Studie gab es Feedback, dass Angaben nicht ganz korrekt seien, dies aber nur mit entsprechendem Vorwissen zu erkennen sei.

H7: Bei weniger bekannten oder verbreiteten Programmen sowie bei Fragen zu Programm-/Betriebsinterna wird ChatGPT Probleme haben *(RQ1, RQ4)*
ChatGPT konnte zu fast allen abgefragten Software Systemen eine Antwort liefern. Bei der nicht verbreiteten *App Mantrailing* verwies ChatGPT darauf, nach einer Hilfefunktion in der App zu schauen. Bei Interna lieferte ChatGPT ebenfalls eine Erklärung oder mindestens eine Vermutung und erklärte, dass dies eine Designentscheidung sei.

H8: Für den Nutzer ist die Korrektheit der Antworten wichtig, von ausschlaggebender Rolle ist jedoch die verständliche, richtungsweisende Qualität *(RQ3)*
Das Feedback diesbezüglich war unterschiedlich; einige Nutzer sahen ChatGPT als gute Ausgangsbasis, während andere Nutzer sich an falschen Informationen sehr störten. In der Interview-Studie führten falsche Informationen neben einem erschwerten Vorankommen dazu, dass der Teilnehmer dachte, dass er das Problem sei und sich dumm fühlte.

H9: Insgesamt wird die Software-Einbindung als hilfreich angesehen *(RQ1, RQ2)*
In der Online-Studie wird die Software-Integration insgesamt als sehr hilfreich angesehen. Es wird jedoch zu bedenken gegeben, dass aufgrund falscher oder nicht angemessener Erklärungen der Nutzen eingeschränkt sein kann. In der Interview-Studie gaben Teilnehmer an, dass eine Einbindung praktisch wäre und ihnen bei der ihrer Arbeit hilfreich sein könnte.

8.3 Limitationen

Im Folgenden wird auf die Limitationen dieser Arbeit nach Wohlin et al. [43] eingegangen. In den dort unterschiedenen Kategorien werden Bedrohungen der Gültigkeit identifiziert und Maßnahmen zur Milderung diskutiert.

Internal Validity *Internal Validity* betrachtet mögliche unbeabsichtigte Einflüsse bzw. Störfaktoren bei der Erhebung der bezweckten Messungen.

In der Online-Studie konnten die Teilnehmer die zu bewertenden Aspekte der Erklärbarkeit, insbesondere die Effektivität und Effizienz, nicht durch eigenes Testen herausfinden oder die Korrektheit der Antwort von ChatGPT überprüfen. Der Nutzer musste sich vorstellen, inwieweit ihm die Erklärungen in der Situation hilfreich sein oder ein schnelleres Arbeiten ermöglichen könnten. Es ist also möglich, dass sich die theoretische Einschätzung von der bei der praktischen Ausführung unterscheidet. Somit besteht das Problem des *hypothetical bias* [6, 22]. Um dem entgegenzuwirken, wurde eine praxisnahe zweite Studie durchgeführt. In beiden Studien wird ChatGPT nicht als Software-Integration als solches betrachtet, da dies den Rahmen dieser Arbeit sprengen würde und außerdem die Machbarkeit sehr eingeschränkt ist.

Allgemein sind die Bewertungen subjektiv begründet; es kann nicht beurteilt werden, welche Kriterien der Teilnehmer zur Bewertung hinzugezogen hat. Somit besteht u.a. keine Sicherheit, dass die Antworten der Teilnehmer ernst gemeint sind. In der Studie hätte eine Kontrollfrage eingebaut werden können, welche zumindest sicherstellt, dass Teilnehmer die Fragen und Antworten lesen. Im Nachhinein wurde versucht, auffällige Muster und Einträge manuell zu identifizieren und herauszufiltern.

Bei der Interview-Studie besteht ein möglicher Einfluss des Interviewers. Die Anwesenheit einer beobachtenden Person kann das Verhalten des Teilnehmers verändert haben. Stellenweise griff die Interviewerin ein, um Hinweise zum Arbeiten mit der generativen KI zu geben. Der Eingriff liegt im Rahmen der Realität und ist mit einem Tutorial zu vergleichen, welches vorab zur Nutzung von ChatGPT gegeben werden könnte.

External Validity *External Validity* bezieht sich auf die Generalisierbarkeit der Ergebnisse.

Die Ausgangsdatenlage für die Studien beschränkt sich auf Alltagssysteme. Dafür wurden mehrere verschiedene Systeme in die Evaluierung einbezogen, sodass die Generalisierbarkeit begünstigt wird. Aufgrund des Fokus auf alltägliche Systeme ist es nicht möglich, die Ergebnisse auf Expertensysteme zu übertragen. Hier könnte anderer Erklärungsbedarf anfallen, sodass ChatGPTs Fähigkeit der Erklärbarkeit in diesem Rahmen gesondert geprüft werden sollte.

Für die Online-Studie wurde ein heterogenes Nutzerspektrum angestrebt; jüngere Erwachsene waren allerdings in der Mehrheit. Daher ist es schwierig, die Ergebnisse auf ältere Personen zu generalisieren. Die gesammelten Daten der ab 40-Jährigen sowie ab 60-Jährigen wiesen jedoch zu denen der unter 40 Jahre alten Probanden keinen signifikanten Unterschied auf.

Die Ergebnisse der Interview-Studie sind auf erfahrene ChatGPT Nutzer nicht anwendbar. Es zeichnete sich während der Umfrage ab, dass mit zunehmender Übung bessere Ergebnisse erzielt wurden. Dies wäre weniger eine Gefährdung, eher eine positive Bestätigung der Anwendbarkeit von ChatGPT.

Insgesamt ist die Güte der Erklärbarkeit schwer zu erfassen. Durch die identifizierten Metriken wurde versucht, verschiedene Ansätze für die Bewertung zu Rate zu ziehen und unterschiedliche Blickwinkel zu betrachten. Gleiches gilt jedoch auch für die Bewertung manuell entwickelter Erklärungen. Ein Vergleich ließe die Gültigkeit der Ergebnisse unter Verwendung der angewandten Metriken leichter einordnen bzw. generalisierbarer machen.

Construct Validity *Construct Validity* erfasst Fehler oder Probleme, welche mit der Experimentplanung in Zusammenhang stehen. In dieser Arbeit betrifft das die Planung der beiden Studien.

In der ersten Studie wurden die Erklärungen von ChatGPT nicht in Echtzeit generiert. Der Grund dafür war, dass eine API-Einbindung mit nicht abschätzbaren Kosten verbunden gewesen ist. Um die Varianz zu erfassen, welche in den Antworten von ChatGPT enthalten ist, wurden jeweils zwei Erklärungen zu einem Prompt generiert und den Teilnehmern zufällig zur Bewertung zugewiesen. Des Weiteren musste der Teilnehmer sich vorstellen, den Erklärungsbedarf eines anderen zu haben und seine Bewertung daraufhin konstruieren. Das begünstigt den *hypothetical bias*.

Um beiden Aspekten entgegenzuwirken, konnten die Teilnehmer in einer zweiten Studie eigene aufkommende Fragen stellen und entsprechend die generierten Antworten aus der eigenen Perspektive bewerten.

Conclusion Validity *Conclusion Validity* betrachtet die Gültigkeit der auf den Ergebnissen basierenden Schlussfolgerungen.

Ein Problem hierbei könnte sein, dass in der Interview-Studie nur eine kleine Stichprobe von Teilnehmern befragt wurde. Zudem konnten für die Studie keine erfahrenen ChatGPT Nutzer akquiriert werden. Daher basieren die Schlussfolgerungen auf einer kleinen Ergebnismenge von noch unerfahrenen Probanden. Die Schlussfolgerungen sind jedoch als Ergänzung zur Hauptuntersuchung zu sehen und berücksichtigen das Vorwissen der Teilnehmer. Zudem ist es kritischer und als Basis anzusehen, gute Erklärbarkeit insbesondere auch für unerfahrene Nutzer einer generativen KI erzeugen lassen zu können.

Kapitel 9

Fazit

In diesem Kapitel werden die wichtigsten Ergebnisse zusammengefasst sowie ein Ausblick auf die Weiterentwicklung und weitere Forschung dieses Ansatzes sowie des Einsatzes von generativer KI gegeben.

9.1 Zusammenfassung

Ziel dieser Arbeit war es, herauszufinden, inwieweit es für Endanwender sinnvoll ist, ChatGPT als Integration in Software Systemen vorzusehen, um Erklärungen zu gewährleisten und Erklärungsbedarf zu befriedigen. Dies ist insbesondere aufgrund komplexer werdender Strukturen der Anwendungssysteme und erweiterten Funktionen für jede Art Hilfestellung ein möglicher Lösungsansatz. Hierzu wurden zwei Nutzerstudien vorbereitet, entwickelt und durchgeführt. Da es im Wesentlichen um die Probleme der Nutzer ging, wurde deren Meinung und Erklärungsbedarf systematisch untersucht.

In dieser Arbeit wird ChatGPT nicht als Software-Integration als solches betrachtet, sondern auf unterschiedliche Art und Weise eine gedachte Anbindung simuliert: Bei der Online-Studie war es durch vorgegebene Fragen und vorgenerierten Antworten nicht erforderlich. Bei der Interview-Studie wurde es durch gleichzeitige Nutzung der vom Nutzer ausgewählten Software sowie parallelem Verwenden von ChatGPT auf zweitem Bildschirm ermöglicht. Der Aspekt der Eignung wird indirekt darüber betrachtet, ob ChatGPT generell in der Lage ist, die Erklärbarkeit eines Systems zu übernehmen. Durch die Online-Studie konnten quantitativ ausreichend Daten gewonnen und durch die Interview-Studie vertiefendere qualitative Nutzereindrücke gesammelt werden.

Die Ergebnisse der Studien sind durchaus positiv; insgesamt erhielten die Erklärungen von ChatGPT eine hohe Zustimmung der Nutzer. Dies gilt im Generellen für Verständlichkeit, Zufriedenheit, Angemessenheit, Effektivität der Antworten sowie denkbar positiver Auswirkung auf die Effizienz. Dabei war ChatGPT in der Lage, für unterschiedliche Erklärungsbedarfe, Erklärungstypen und Anwender zu reagieren. Zum Zwecke der Objektivierung wurden Lesbarkeitsindizes betrachtet. Als Schwachstellen erwiesen sich bei ChatGPT die Korrektheit der Antworten und die Länge der Antworten, insofern keine sehr konkretisierte Begrenzung voreingestellt wurde. Kontraproduktiv verhält sich bei einer Begrenzung aber die Güte der Warum-Erklärungen. Hier ist also ein noch zu lösender Konflikt absehbar.

Insgesamt ist die in dieser Arbeit untersuchte und diskutierte Anwenderunterstützung durch Einbindung generativer KI zu empfehlen.

9.2 Ausblick

In dieser Arbeit wurden erste Untersuchungen zum Einsatz von generativer KI zur Generierung von Erklärungen für Software Systeme vorgenommen. Diese könnten in weiterer Forschung näher betrachtet und ausgebaut werden.

Interessant wäre ein Vergleich zwischen der Nutzung mit und ohne ChatGPT auf Basis von Nutzergruppen mit identischer Ausgangsbasis. Dies ist in der Realität jedoch schwierig umzusetzen, da derart Probanden kaum zu finden sind. Dennoch könnte zumindest Gruppen mit ähnlicher Ausgangsbasis auf den Einfluss der Art der Erklärungsbereitstellung untersucht werden. Einfacher zu realisieren ist hingegen, formulierte Erklärungen von Entwicklern mit den KI-Generierten zu vergleichen.

Eine genauere Anpassung von ChatGPT durch das Verwenden von individualisiterten Modellen ist ebenfalls ein weiterer Ansatz zur Forschung. Hier könnte die Art des Erreichens der „richtigen“ Antwortengebung und entsprechend angepasster Form der Ausgabe zu untersuchen sein. Außerdem interessant wäre hier eine Anpassung an die Software System und den Nutzer, beispielsweise durch eine grobe Erhebung des Nutzerwissenstands zu Beginn der Interaktion mit ChatGPT. In diesem Rahmen ist auch der Einfluss von bereitgestellten Hinweisen zum Arbeiten mit der generativen KI ein interessanter Untersuchungspunkt.

Generative KI ist in dieser Arbeit insbesondere im Zusammenhang mit der Hilfestellungen bei Erklärungsbedarf betrachtet worden. Eine Einbindung ist jedoch in allen denkbar möglichen Anwendungsfeldern vorstellbar; und mit einer Weiterentwicklung sowohl von den Möglichkeiten als auch der generativen KI selbst mit Sicherheit in Zukunft zu rechnen.

Erste Systeme fangen bereits an, KI zu integrieren; nicht nur im Allgemeinen als Unterstützungshilfe, wie zur Generierung von Text- und Layout-Vorschläge in Notion, sondern auch konkret zur Beantwortung von Fragen. Beispielhaft sei der Foxit PDF Reader AI Assistant genannt, welcher den hier in der Arbeit untersuchten Ansatz der Integration von ChatGPT bereits beginnt umzusetzen [1]. Diese Funktion ist mit eingeschränkter Nutzung und entsprechenden Haftungsausschlüssen seit Ende 2023 verfügbar. Es wird dabei deutlich darauf hingewiesen, dass ungenaue Informationen über Personen, Orte oder Fakten geliefert werden könnten; also der gesunde Menschenverstand durchaus noch gefragt ist.

[1] https://kb.foxit.com/hc/en-us/articles/22886102196116-Foxit-PDF-Editor-and-Foxit-PDF-Reader-has-integrated-with-ChatGPT; letzter Aufruf: 22.04.2024

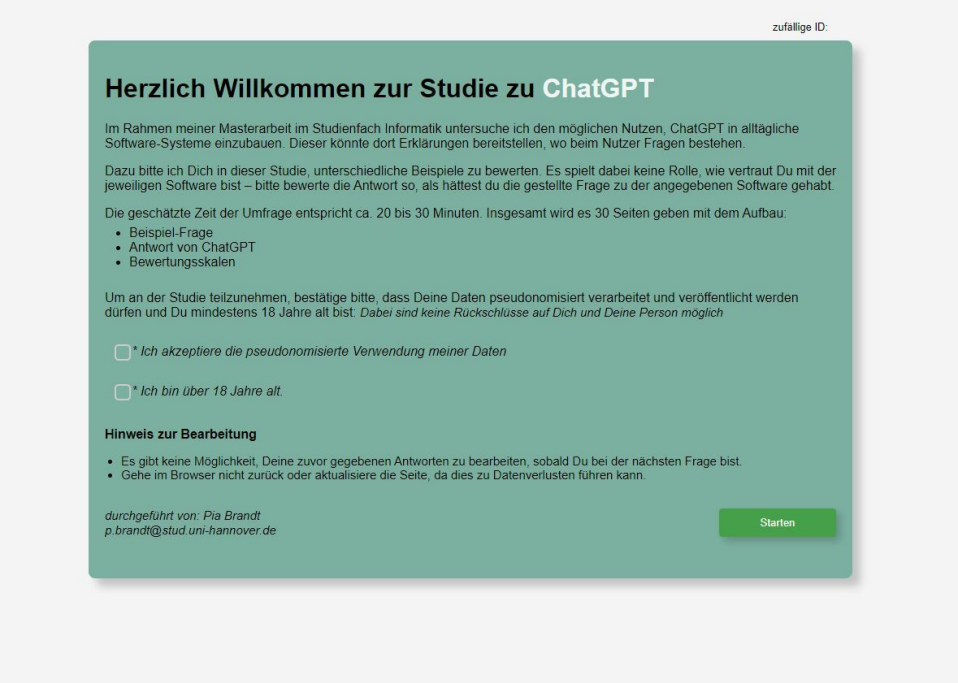
Anhang A

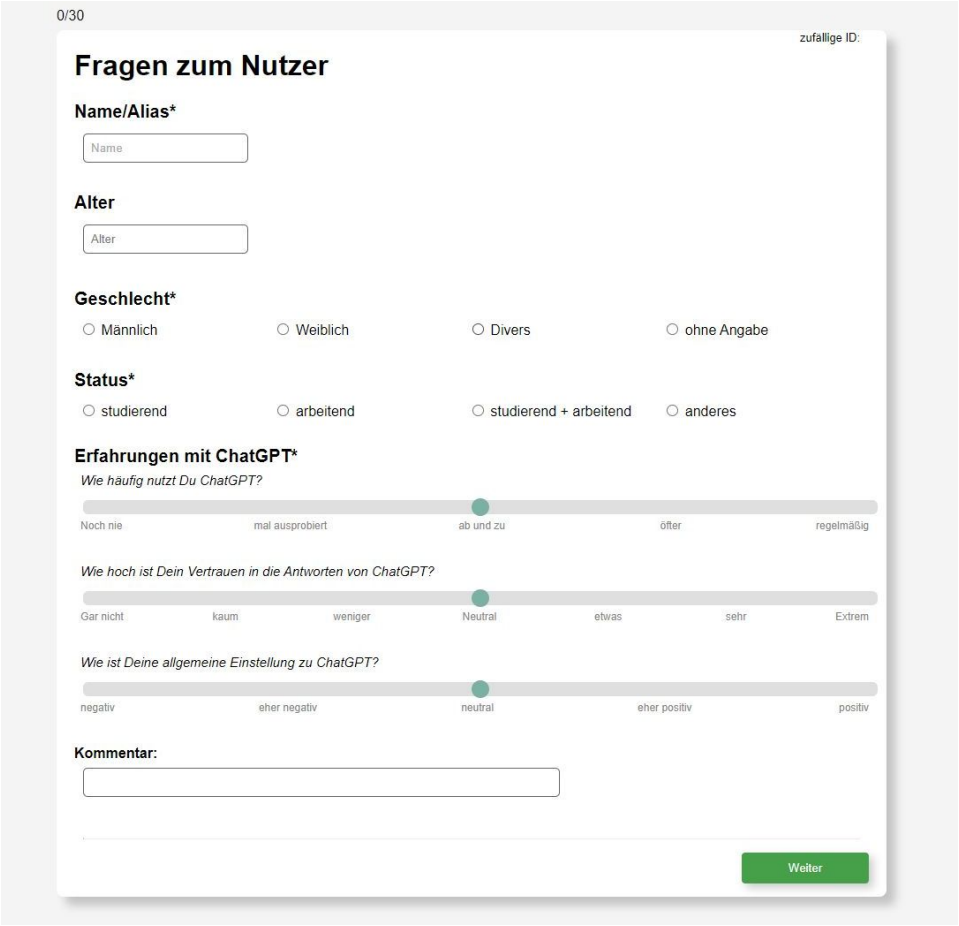
Online-Studie

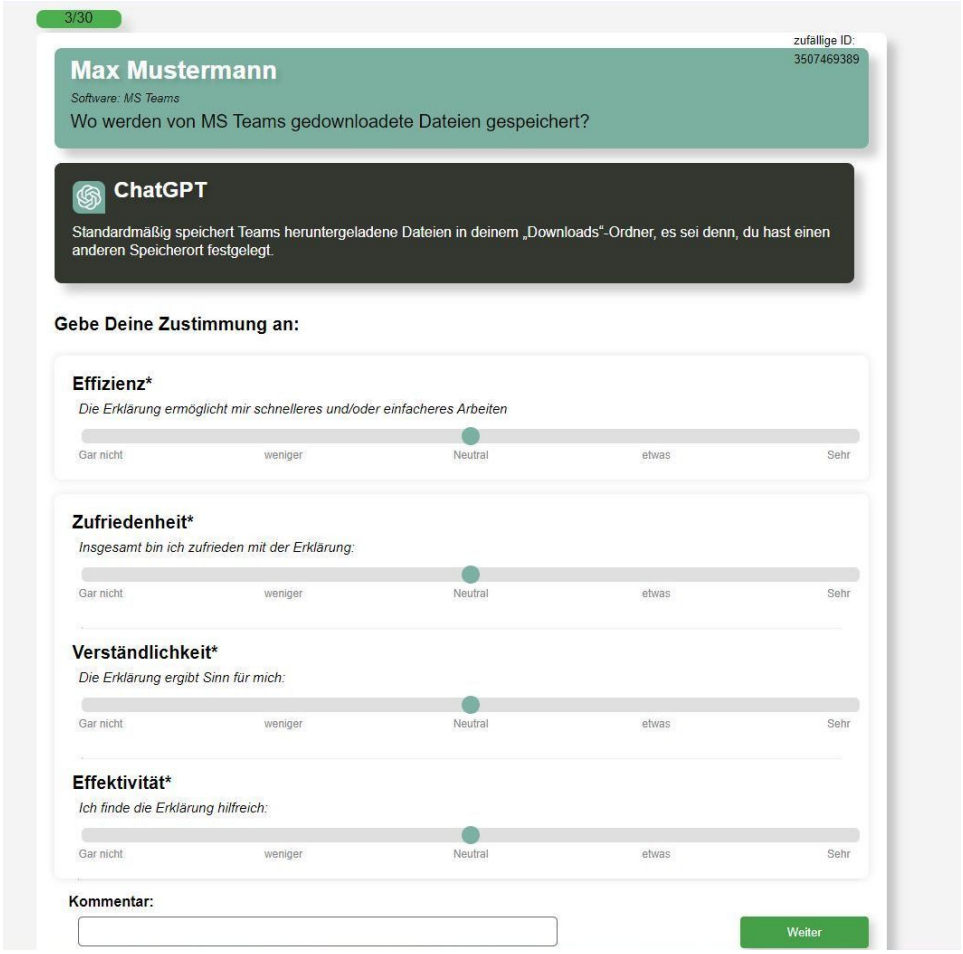
Im Folgenden werden die Anhänge zur Online-Studie abgebildet. Sie enthalten beispielhafte Screenshots der Seiten der Online-Studie. Des Weiteren werden die in der Umfrage vorkommenden Software Systeme aufgelistet, zu denen die Teilnehmer zufällig Prompts und Erklärungen erhielten.

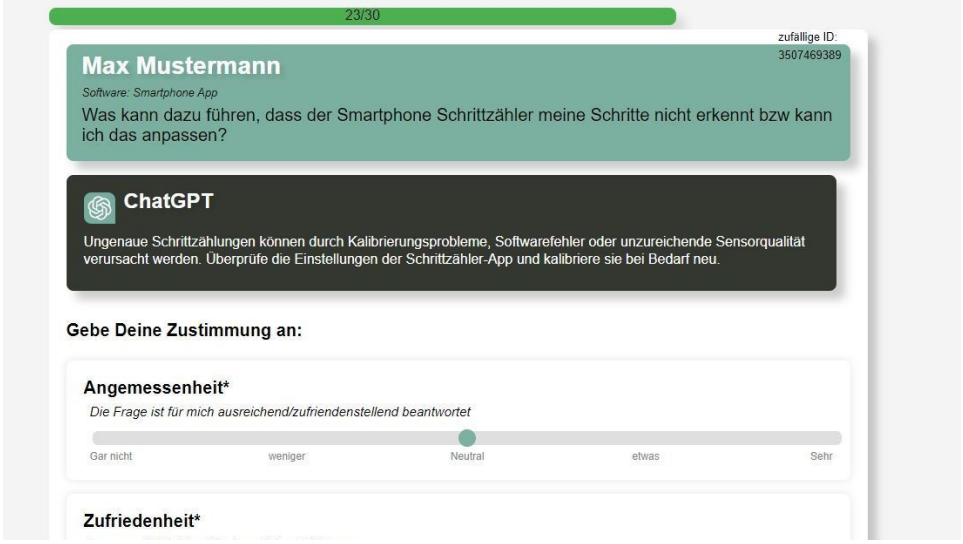
A.1 Screenshots der Online-Studie

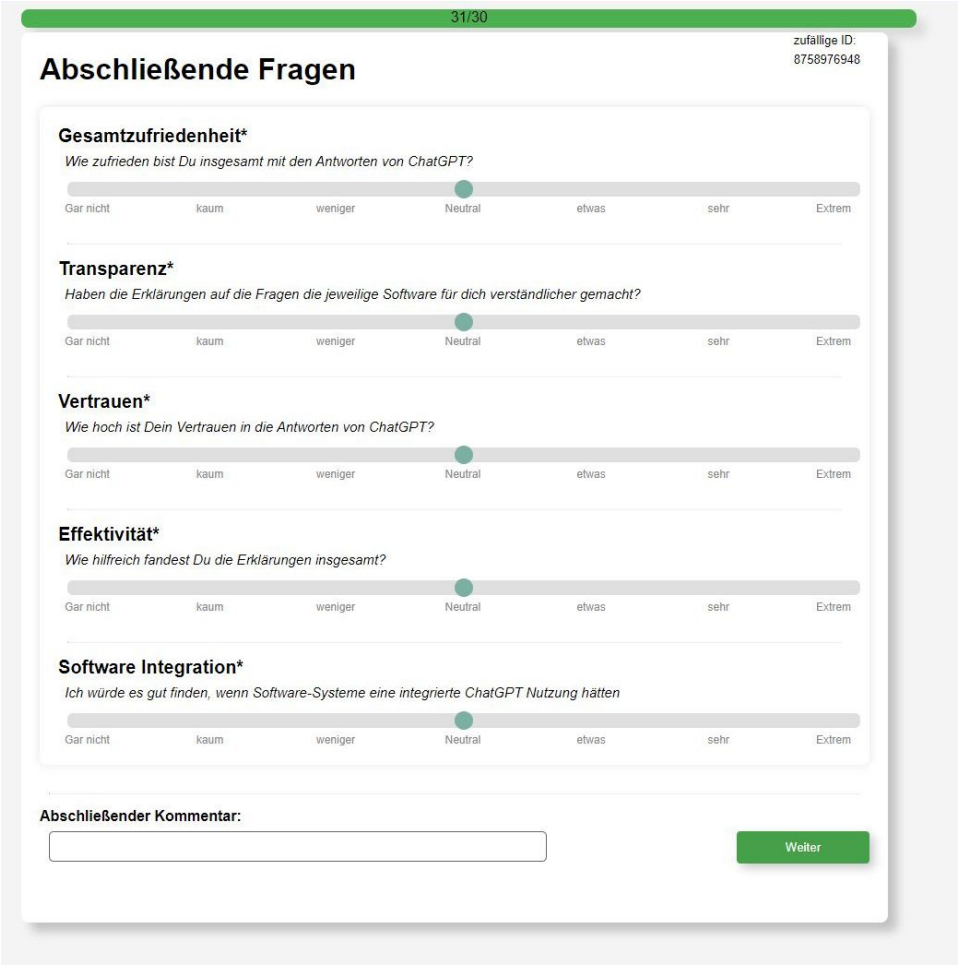
Die Beispiele enthalten neben dem Studienrahmen zwei Frage-Antwort-Seiten inklusive den kategoriespezifischen Fragen zu Effizienz und Angemessenheit.




0/30
zufällige ID:
Fragen zum Nutzer
Name/Alias*
Name
Alter
Alter
Geschlecht*
Männlich
Weiblich
Divers
ohne Angabe
Status*
studierend
arbeitend
studierend + arbeitend
anderes
Erfahrungen mit ChatGPT*
Wie häufig nutzt Du ChatGPT?
Noch nie
mal ausprobiert
ab und zu
öfter
regelmäßig
Wie hoch ist Dein Vertrauen in die Antworten von ChatGPT?
Gar nicht
kaum
weniger
Neutral
etwas
sehr
Extrem
Wie ist Deine allgemeine Einstellung zu ChatGPT?
negativ
eher negativ
neutral
eher positiv
positiv
Kommentar:
Weiter


3/30
zufällige ID:
3507469389
Max Mustermann
Software: MS Teams
Wo werden von MS Teams gedownloadete Dateien gespeichert?
ChatGPT
Standardmäßig speichert Teams heruntergeladene Dateien in deinem „Downloads"-Ordner, es sei denn, du hast einen anderen Speicherort festgelegt.
Gebe Deine Zustimmung an:
Effizienz*
Die Erklärung ermöglicht mir schnelleres und/oder einfacheres Arbeiten
Gar nicht
weniger
Neutral
etwas
Sehr
Zufriedenheit*
Insgesamt bin ich zufrieden mit der Erklärung:
Gar nicht
weniger
Neutral
etwas
Sehr
Verständlichkeit*
Die Erklärung ergibt Sinn für mich:
Gar nicht
weniger
Neutral
etwas
Sehr
Effektivität*
Ich finde die Erklärung hilfreich:
Gar nicht
weniger
Neutral
etwas
Sehr
Kommentar:
Weiter


23/30
zufällige ID:
3507469389
Max Mustermann
Software: Smartphone App
Was kann dazu führen, dass der Smartphone Schrittzähler meine Schritte nicht erkennt bzw kann ich das anpassen?
ChatGPT
Ungenaue Schrittzählungen können durch Kalibrierungsprobleme, Softwarefehler oder unzureichende Sensorqualität verursacht werden. Überprüfe die Einstellungen der Schrittzähler-App und kalibriere sie bei Bedarf neu.
Gebe Deine Zustimmung an:
Angemessenheit*
Die Frage ist für mich ausreichend/zufriendenstellend beantwortet
Gar nicht
weniger
Neutral
etwas
Sehr
Zufriedenheit*


31/30
zufällige ID:
8758976948
Abschließende Fragen
Gesamtzufriedenheit*
Wie zufrieden bist Du insgesamt mit den Antworten von ChatGPT?
Gar nicht kaum weniger Neutral etwas sehr Extrem
Transparenz*
Haben die Erklärungen auf die Fragen die jeweilige Software für dich verständlicher gemacht?
Gar nicht kaum weniger Neutral etwas sehr Extrem
Vertrauen*
Wie hoch ist Dein Vertrauen in die Antworten von ChatGPT?
Gar nicht kaum weniger Neutral etwas sehr Extrem
Effektivität*
Wie hilfreich fandest Du die Erklärungen insgesamt?
Gar nicht kaum weniger Neutral etwas sehr Extrem
Software Integration*
Ich würde es gut finden, wenn Software-Systeme eine integrierte ChatGPT Nutzung hätten
Gar nicht kaum weniger Neutral etwas sehr Extrem
Abschließender Kommentar:
Weiter


Vielen Dank für die Teilnahme!
zufällige ID:
8758976948
Die Antworten wurden erfolgreich gespeichert!
(Die Seite kann nun geschlossen werden)

A.2 Software Systeme

In der Studie wurden Fragen und Antworten zu folgenden Software Systemen präsentiert:

- Aktionen.consors-bank.de
- AutoCAD
- Avira
- BlueJ
- Chirp
- Chrome
- Discord
- Disney+
- Eclipse
- Facebook
- Firefox
- Goodnotes
- Google
- Herd
- Instagram
- iTerm2
- League of Legends
- LinkedIn
- MS Azure
- MS Excel
- MS Outlook
- MS Powerpoint
- MS Teams
- MS Word
- Netflix
- Notion
- Online Banking
- Opera Gx Web-browser
- Reddit
- Smartphone App
- Spotify
- SQL Server Management Studio (SSMS)
- StudIP
- Telegram
- Thunderbird
- Tiktok
- Udemy.com
- Valorant Game
- WhatsApp
- Windows
- Youtube
- Zoom

Anhang B

Interview-Studie

Zur Durchführung der Interview-Studie wurde ein Leitfaden für den Interviewer sowie eine Version für den Teilnehmer erstellt.

B.1 Leitfaden für den Teilnehmer

Nutzerstudie:

Live-Session mit ChatGPT und Software Fragen

1. Teil: Einführung zur Studie

In dieser Studie wird untersucht, inwieweit ChatGPT den Arbeitsprozess unterstützen bzw. verbessern und bei Unklarheiten oder Fragen zu Software Systemen hilfreiche Antworten liefern kann. Du wirst dabei live mit ChatGPT interagieren, eigene Anfragen stellen und ein bisschen herumprobieren können.

Insgesamt wird die Studie ca. 30 bis 60 Minuten dauern und in einer von Dir gewählten, bekannten Arbeitsumgebung oder am SE Institut in der Leibniz Universität Hannover stattfinden.

Zunächst werden ein paar Angaben zu Dir festgehalten werden.

Anschließend wirst Du die Möglichkeit haben, eigene Fragen zu einem beliebigen Software-System Deiner Wahl von ChatGPT klären zu lassen, sofern Du generell welche hast oder Dir im Vorhinein überlegt oder festgehalten hast.

Danach bekommst Du eine Aufgabe je nach Präferenz zu der Software Excel oder Notion, welche Du mit der Hilfe von ChatGPT lösen sollst.

Zwischendurch sowie am Ende werden wir festhalten, wie Deine Meinung zu den Antworten ist, welche Probleme es gibt oder was Dir ansonsten auf- und einfällt.

Du kannst ChatGPT alles fragen - es gibt keine dummen Anfragen!

Mit der Teilnahme stimmst Du der pseudonymisierten Verarbeitung und Veröffentlichung Deiner Daten zu.

➢ Grundlegende Meinung zu ChatGPT

- Was ist deine Meinung/Einstellung zu ChatGPT?
- Was weißt du über ChatGPT?
- Wie hoch ist dein Vertrauen?
- Wie und wie oft nutzt du ChatGPT?
- Hast du Erwartungen an ChatGPT (für diese Studie/diesen Kontext)

2. Teil: Eigene Fragen des Teilnehmers

Hast Du eigene Fragen zu einer Software?

3. Teil: Aufgabe

➢ Notion: Organisation von Daten

Erstelle in Notion (die Struktur für) eine oder mehrere Rezeptsammlungen mit Back- und Kochrezepten nach deinen Vorstellungen.

Führe Anpassungen durch wie:

- Füge mindestens ein Back- und ein Kochrezept hinzu
- (Entferne den Kommentarbereich)
- Erstelle Kategorien wie „Kochen“ und „Backen“ und weise den Rezepten die entsprechende Kategorie zu
- Erstelle einen Abschnitt, in welchem nur Back Rezepte zu sehen sind
- Erstelle einen Abschnitt, in dem alle Rezepte aufgelistet sind
- Erstelle einen Abschnitt, in dem Bilder mit eingebunden werden können und füge ein Beispielbild ein
- Verschönere die ganze Seite

➢ Excel: Verarbeitung von Daten

Erstelle eine Excel-Tabelle, in der Name, Kategorie sowie Kosten von ausgedachten Anschaffungen beinhaltet sind (min. 50 Einträge).
Lasse dir die Gesamtkosten für alle Kategorien berechnen.

Führe Aktionen durch wie:

- Füge mindestens einen sehr teuren Einkauf hinzu
- Bestimme die Summe von zwei Kategorien deiner Wahl
- Führe eine Abschreibung (linear + degressiv) von dem teuren Einkauf durch *(als Tabelle, benötigte Werte können ausgedacht werden)*
- Erstelle eine Grafik, in welcher der Wert (mit beiden Verfahren) des Gegenstandes über die Zeit angezeigt wird
- Anpassung des Aussehens der Grafik

4. Teil: Abschluss

- Wurden deine Erwartungen erfüllt?
- Hat sich etwas über deine Meinung zu ChatGPT geändert?
- Würdest du es weiter empfehlen?
- Würdest du es als Softwareerweiterungen nutzen wollen? Warum?
- Abschließende Meinung:

B.2 Leitfaden der Interview-Studie

Interview Template:

Live-Session mit ChatGPT und Software Fragen

Setup:

- ChatGPT Login, ggf. Notion Login; beide Programme wenn möglich nebeneinander öffnen
- ChatGPT neuer Chat mit Alias, Datum, Software benennen
- Mögliche Präferenzen zu ChatGPT Antworten abklären und einstellen

Aktuelle Anpassung:
answers should be precise but short. If it is something you generally can say more about, its ok for the answer to be longer. ChatGPT can ask questions back to get more into detail. If the question is asking for a way to do something the answer should be a guide to do so. The description of the step should contain not only what to do but also how to do it

Template: Studie mit am

1. Kontext: Wo findet Studie statt, wie ist der Arbeitsplatz?

Ort: Leibniz Universität Hannover/heimischer Arbeitsplatz
Kontext: gestellter Platz für die Studie/gewohnte Arbeitsumgebung
Aufbau: Laptop des Teilnehmers, 2. Bildschirm, PC

2. Teilnehmer Start: ~ 10 Minuten

Angaben zum Teilnehmer

Name/Alias:
Alter:

Geschlecht:	m	w	d	keine Angabe
Status:	Studierend	arbeitend	beides	anderes

Grundlegende Meinung zu ChatGPT

- Was ist deine Meinung/Einstellung zu ChatGPT?
- Was weißt du über ChatGPT?
- Wie hoch ist dein Vertrauen?
- Wie und wie oft nutzt du ChatGPT?
- Hast du Erwartungen an ChatGPT (für diese Studie/diesen Kontext)

3. Eigene Fragen des Teilnehmers **Start:** **0/~ 10 Minuten**4. Inkorrekte Angaben:

- Fragenumformulierungen:
  - wegen ChatGPT:
  - eigener Ungenauigkeit:
- Probleme bei der Nutzung:
- Beobachtungen/Kommentare:

_____

- Zufriedenheit mit Antwort:
- Hast du das Gefühl, es war schneller als herumzuprobieren oder zu googlen?
- Hat es dir die Arbeit erleichtert?
- Hat es dich abgelenkt?
- Hast du etwas über die SW gelernt?
- Fandest du es insgesamt hilfreich?

5. Gestellte Aufgabe **Start: ~ 20-40 Minuten**

Grund der Wahl:

_____

- Inkorrekte Angaben:
- Fragenumformulierungen:
  - wegen ChatGPT
  - eigener Ungenauigkeit:

- Probleme bei der Nutzung:

- Beobachtungen/Kommentare:

_____

- Zufriedenheit mit Antwort:

- Hast du das Gefühl, es war schneller als herumzuprobieren oder zu googlen?

- Hat es dir die Arbeit erleichtert?

- Hat es dich abgelenkt?

- Hast du etwas über die SW gelernt?

- Fandest du es insgesamt hilfreich?

6. Abschluss Start: ~ 10 Minuten

- Wurden deine Erwartungen erfüllt?

- Hat sich etwas über deine Meinung zu ChatGPT geändert?

- Würdest du es weiter empfehlen?

- Würdest du es als Softwareerweiterungen nutzen wollen? Warum?

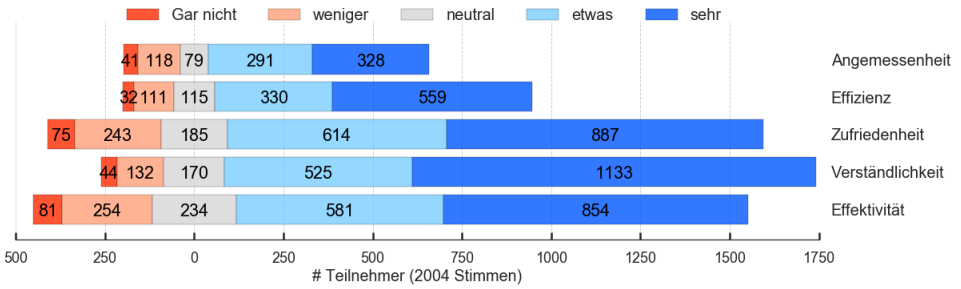
- Abschließende Meinung:

Ende:

Anhang C

Weitere Grafiken zur Auswertung

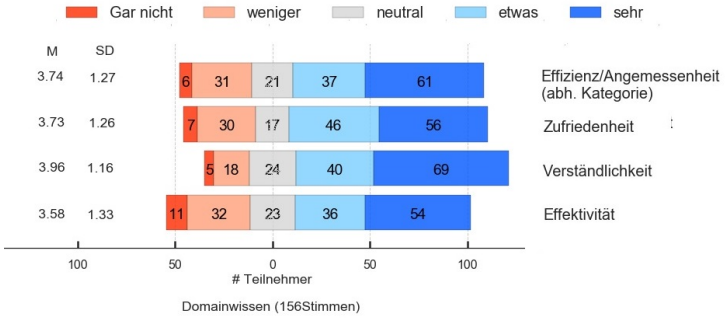
Gesamtbewertung über alle Teilnehmer und Bewertungen zu den Antworten, unterteilte kategoriespezifische Frage: Angemessenheit und Effizienz.



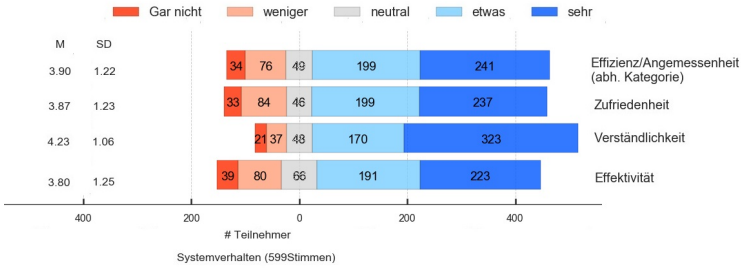
M:
Verständlichkeit 4.28
Effizienz 4.11
Zufriedenheit 4.00
Effektivität 3.93
Angemessenheit 3.87

SD:
Angemessenheit 1.20
Effektivität 1.19
Zufriedenheit 1.17
Effizienz 1.10
Verständlichkeit 1.02

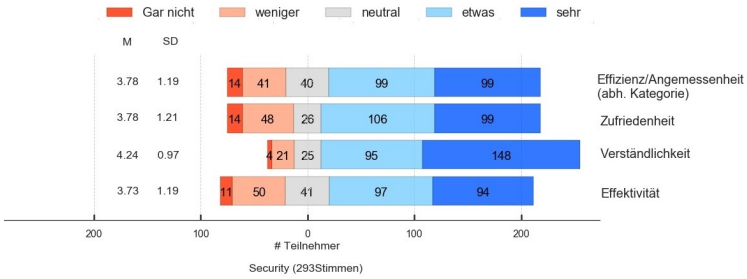
Art von Erklärungsbedarf: Domainwissen



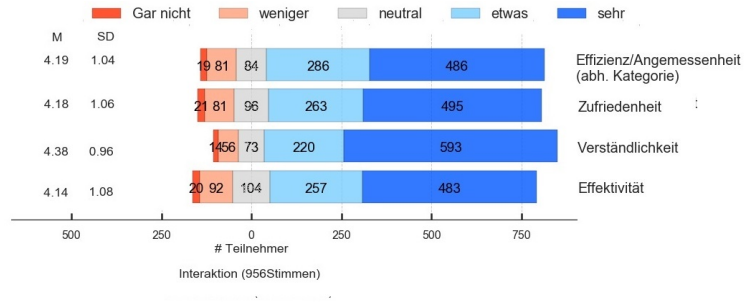
Systemverhalten



Security



Interaktion



Abbildungsverzeichnis

3.1 Vorgehen der Arbeit 17
3.2 Wichtige identifizierte Aspekte zur Bewertung der Erklärbarkeit [12] im Anwendungsszenario 19
3.3 ausgewählte Kriterien und Metriken Zufriedenheit 20
3.4 Ausgewählte Kriterien und Metriken Verständlichkeit 20
3.5 Ausgewählte Kriterien und Metriken Effektivität 21
3.6 Ausgewählte Kriterien und Metriken Korrektheit 22
3.7 Ausgewählte Kriterien und Metriken Effizienz 23
3.8 Ausgewählte Kriterien und Metriken Angemessenheit 23

4.1 Iterationen zur Entwicklung der finalen Daten 30

5.1 Ablauf der Online-Studie 37
5.2 Bewertung der Zustimmung: Ausschnitt einer Frage-Antwort-Seite der Online-Studie 38
5.3 Erfahrungen mit ChatGPT 43
5.4 Höhe des Vertrauens in die Antworten von ChatGPT 43
5.5 Bewertung über alle Erklärungen über alle Teilnehmer 44
5.6 Erklärungen zu Prompt 65 mit der geringsten Zustimmung: *Was bedeuten die Kürzel in den Einstellungen bei Chirp?* . . 46
5.7 Bewertung der Erklärungen zu Prompt 32 Antwort 1 (links) und Antwort 2 (rechts) 47
5.8 Zustimmung der Bewertungen nach Erklärungstyp, Wie-Erklärung (oben), Was-Erklärung (unten links), Warum-Erklärung (unten rechts) 48
5.9 Abschlussbewertungen zu ChatGPT 50

6.1 Ablauf der Interview-Studie 53

Tabellenverzeichnis

4.1 Beschreibung der Labels für das Behalten von Daten (oben) sowie der ausführenden Begründung für die Einordnung (unten) 31

5.1 Vergleich der Bewertungen nach Alter und Geschlecht 45
5.2 Vergleich der beiden generierten Antworten 45

7.1 Mittelwerte Sätze und Verben der beiden generierten Erklärungen (*nur für Prompts aus der Online-Studie) 67

Literaturverzeichnis

[1] T. Amstad. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service, 1978.

[2] D. Baidoo-anu and L. Owusu Ansah. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, 7(1):52–62, 2023.

[3] W. BRIEST. Kann man verständlichkeit messen? *STUF - Language Typology and Universals*, 27(1-3):543–563, 1974.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[5] E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. Number 31161 in Working Paper Series. National Bureau of Economic Research, April 2023.

[6] J. Buckell, J. Buchanan, S. Wordsworth, F. Becker, L. Morrell, A. Roope, L.and Kaur, and L. Abel. Hypothetical bias. *Catalogue of Bias*, 2020.

[7] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

[8] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: A definition, a model, and a knowledge catalogue. In *2021 IEEE 29th*

International Requirements Engineering Conference (RE), pages 197–208, 2021.

[9] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25:493–514, 2020.

[10] D. R. Cotton, P. A. Cotton, and J. R. Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, 61(2):228–239, 2024.

[11] M. Debelak. Entwicklung von kriterien zur bewertung von erklärungen im rechnungswesenunterricht: eine analyse der einschätzung von studienanfängerinnen der wirtschaftspädagogik an den universitäten graz, wien und innsbruck. Universitäten Graz, 2017.

[12] H. Deters, J. Droste, M. Obaidi, and K. Schneider. How explainable is your system? towards a quality model for explainability. In *Proceedings of the 30th International Working Conference on Requirement Engineering: Foundation for Software Quality*. Springer, 2024.

[13] H. Deters, J. Droste, and K. Schneider. A means to what end? evaluating the explainability of software systems using goal-oriented heuristics. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, page 329–338, New York, NY, USA, 2023. Association for Computing Machinery.

[14] J. H. S. Dongmo, M. Krüßmann, and F. Weimann. Chatgpt–dein freund und helfer im hochschulalltag? FH Münster, 2023.

[15] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. Cornell University, 2017.

[16] J. Droste, H. Deters, M. Obaidi, and K. Schneider. Explanations in everyday software systems: Towards a taxonomy for explainability needs. In *2024 IEEE 32nd international requirements engineering conference (RE)*. IEEE, 2024.

[17] J. Euchner. Generative ai. *Research-Technology Management*, 66(3):71–74, 2023.

[18] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[19] K. Fuchs. Exploring the opportunities and challenges of nlp models in higher education: is chat gpt a blessing or a curse? *Frontiers in Education*, 8, 2023.

[20] A. Geist. Rechtsinformation, rechtsdatenbanken & chatgpt: Eine erste einordnung. *Jusletter IT*, 29, 2023.

[21] R. Gunning. The technique of clear writing. McGraw-Hill, 1952.

[22] D. A. Hensher. Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological*, 44(6):735–752, 2010. Methodological Advancements in Constructing Designs and Understanding Respondent Behaviour Related to Stated Preference Experiments.

[23] K.-A. Immel. *Regionalnachrichten im Hörfunk - Verständlich schreiben für Radiohörer*. Springer VS Wiesbaden, 2014.

[24] S. Kabir, D. N. Udo-Imeh, B. Kou, and T. Zhang. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions. Cornell University, 2024.

[25] K. S. Kalyan. A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, 6:100048, 2024.

[26] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[27] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Institute for Simulation and Training, University of Central Florida, 1975.

[28] J. Klein. Erklären-was, erklären-wie, erklären-warum. typologie und komplexität zentraler akte der welterschließung. *Erklären. Gesprächsanalytische und fachdidaktische Perspektiven*, 2:25–36, 2009.

[29] C. Krech. *Erklärbarkeit als Schlüssel für den verantwortungsvollen Umgang mit KI*, pages 83–117. Springer Fachmedien Wiesbaden, Wiesbaden, 2023.

[30] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368. IEEE, 2019.

[31] T.-C. Lee, K. Staller, V. Botoman, M. P. Pathipati, S. Varma, and B. Kuo. Chatgpt answers common patient questions about colonoscopy. *Gastroenterology*, 165(2):509–511.e7, 2023.

[32] W. M. Lim, A. Gunasekara, J. L. Pallant, J. I. Pallant, and E. Pechenkina. Generative ai and the future of education: Ragnarök or reformation? a paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2):100790, 2023.

[33] B. D. Lund and T. Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 40(3):26–29, 2023.

[34] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis. Xai for all: Can large language models simplify explainable ai? *arXiv preprint arXiv:2401.13110*, 2024.

[35] R. Michel-Villarreal, E. Vilalta-Perdomo, D. E. Salinas-Navarro, R. Thierry-Aguilera, and F. S. Gerardou. Challenges and opportunities of generative ai for higher education as explained by chatgpt. *Education Sciences*, 13(9), 2023.

[36] S. Naumann, A. Guldner, S. Weber, and M. Westing. Was weiß chatgpt über nachhaltige software-entwicklung und green coding? erste tests und bewertungen. In *INFORMATIK 2023 - Designing Futures: Zukünfte gestalten*, pages 1277–1288. Gesellschaft für Informatik e.V., Bonn, 2023.

[37] J. V. Pavlik. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1):84–93, 2023.

[38] L. J. Quintans-Júnior, R. Q. Gurgel, A. A. d. S. Araújo, D. Correia, and P. R. Martins-Filho. Chatgpt: the new panacea of the academic world. *Revista da Sociedade Brasileira de Medicina Tropical*, 56:e0060–2023, 2023.

[39] A. S. Rahat Khan, Nidhi Gupta and R. Chakravarty. Impact of conversational and generative ai systems on libraries: A use case large language model (llm). *Science & Technology Libraries*, pages 1–15, 2023.

[40] T. Speith and M. Langer. A new perspective on evaluation methods for explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 325–331. IEEE, 2023.

[41] W. Stegmüller. *Das ABC der modernen Logik und Semantik: Der Begriff der Erklärung und seine Spielarten*. Springer, 1969.

[42] F. Witte. *Metriken für Usability-Tests*, pages 173–178. Springer Fachmedien Wiesbaden, Wiesbaden, 2018.

[43] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[44] G. Yenduri, R. M, C. S. G, S. Y, G. Srivastava, P. K. R. Maddikunta, D. R. G, R. H. Jhaveri, P. B, W. Wang, A. V. Vasilakos, and T. R. Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv preprint arXiv:2305.10435, 2023.