Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Text oder Sprache?
Ein Vergleich verschiedener Interaktionsmodalitäten im Kontext der Informationsbeschaffung

Bachelorarbeit im Studiengang Mensch-Computer-Interaktion

Autor:
Leon Korkmaz

Matrikel: 0000000

Human-Computer Interaction
Fachbereich Informatik

Erstgutachter:	Prof. Dr. Frank Steinicke
Zweitgutachter:	Dr. Susanne Schmidt
Betreuer:	Dr. Susanne Schmidt

Hamburg, 28. Februar 2024

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

Text oder Sprache? Ein Vergleich verschiedener Interaktionsmodalitäten im Kontext der Informationsbeschaffung

Leon Korkmaz*
Universität Hamburg

ZUSAMMENFASSUNG

Die rasche Entwicklung künstlicher Intelligenzen macht es erforderlich, geeignete Modalitäten für eine effiziente und benutzerfreundliche Interaktion zu finden. Die Interaktion kann sowohl über Text als auch über Sprache erfolgen. Während Chatbots und Sprachassistenten jeweils auf eine Modalität setzen, werden in dieser Arbeit verschiedene Modalitätskombinationen miteinander verglichen. Dazu wurde eine Anwendung mit verschiedenen Ein- und Ausgabemodalitäten zur Informationssuche entwickelt und die Kombinationen mithilfe einer 2x2 Within-Subject Benutzerstudie evaluiert. Es zeigt sich, dass effiziente Modalitätskombinationen nicht immer bevorzugt werden und dass Usability Faktoren, wie die Benutzerkontrolle, einen größeren Einfluss auf die Präferenz haben. So kann sich die Kombination von Spracheingabe und Textausgabe als sehr effizient erweisen, gleichzeitig aber durch Schwächen in der User Experience weniger präferiert werden. Die rein textbasierte Interaktion setzt sich als bevorzugte Kombination durch, da sie den Nutzenden ein hohes Maß an Kontrolle und Freiheit gewährleistet. Dialogbasierte Anwendungen zur Informationssuche sollten daher die Modalität Text als Basis nutzen und diese durch sprachbasierte Ein- und Ausgabemöglichkeiten ergänzen.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Text input; Human-centered computing—Human computer interaction (HCI)—Interaction devices—Sound-based input / output Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Natural language interfaces

1 EINLEITUNG

Mit dem Aufkommen fortgeschrittener Sprachmodelle wie ChatGPT [18], stellt sich die Frage, wie wir in Zukunft mit künstlichen Intelligenzen (KI) interagieren werden. Das Generative Pre-trained Transformer (GPT) Modell von OpenAI verfügt bereits heute über die Fähigkeit, menschliche Sprache zu verstehen und Antworten auf eine sehr natürliche Art und Weise [10] zu generieren. Durch die vielfältigen Einsatzmöglichkeiten sowohl im privaten Kontext als auch in der Wirtschaft und der Forschung, werden sich diese KI-Systeme zunehmend in unseren Alltag integrieren. Umso wichtiger ist es, die Interaktion mit solchen Systemen effizient und benutzerfreundlich zu gestalten.

Das Forschungsfeld Human-Computer Dialogue (HCD) untersucht seit dem Aufkommen Künstlicher Intelligenzen die Interaktion zwischen Mensch und Maschine [1]. Maschinen, mit denen über eine natürlichsprachliche Schnittstelle interagiert werden kann, werden als HCD-Systeme [33], Dialogsysteme oder auch als Conversational Agents bezeichnet. Die Interaktion kann über verschiedene Modalitäten erfolgen, meist wird auf Text oder Sprache zurückgegriffen. Text als Interaktionsmodalität wird häufig im Bereich von Chatbots im Kundensupport [28] eingesetzt. Die Interaktion mit ChatGPT über die von OpenAI bereitgestellte Benutzeroberfläche erfolgt derzeit ebenfalls im Stil eines Chatbots mit Texteingabe und Textausgabe. In den letzten Jahren hat sich die Sprache als natürlichere und intuitivere Modalität [28] im Vergleich zu Text etabliert. Der enorme Zuwachs an Rechenleistung [14], sowie der Einsatz von Deep Neural Networks und neuen Algorithmen des maschinellen Lernens haben zu einer stetigen Verbesserung der Spracherkennung und Sprachsynthese geführt [13]. So hat die sprachbasierte Interaktion ihren Einzug in virtuelle Assistenzsysteme wie Siri erhalten [14]. Trotz dieser stetigen Weiterentwicklung sprachbasierter Systeme ist es schwierig, eine Modalität als überlegen anzusehen [19]. Die Wahl der richtigen Modalität muss in Abhängigkeit von den Charakteristiken des Nutzenden und der Aufgabe getroffen werden [19]. Dialogsysteme erfüllen unterschiedliche Aufgaben und können in zielgerichtete und nicht-zielgerichtete Systeme klassifiziert werden [33]. Zielgerichtete Systeme sind darauf ausgelegt, bestimmte Informationen zu finden oder einfache Aufgaben auszuführen. Ein Beispiel sind Assistenzsysteme, die Nutzende bei der Buchung von Tickets unterstützen. Nicht-zielgerichtete Systeme können mit Menschen interagieren, ohne ein bestimmtes Ziel zu verfolgen. Sie erlauben eine freie, explorative Interaktion mit dem System. Nicht zielgerichtete Funktionen finden sich daher vor allem in Chatbots und virtuellen Agenten, die eine freie Konversation zwischen Mensch und Maschine ermöglichen. Die Grenze zwischen den Systemen ist nicht strikt definiert, oft finden sich Systeme, die sowohl zielgerichtete als auch nicht-zielgerichtete Dialoge führen können [33]. Als Beispiel kann ein virtueller Agent wie Siri dienen, der sowohl einen Chat anbietet und damit eine nicht zielgerichtete Konversation aufrechterhält, als auch zielgerichtete Antworten auf gestellte Fragen liefert und Aufgaben automatisieren kann. Neben der Aufgabenstellung kann auch der übergeordnete Kontext eine Rolle bei der Auswahl geeigneter Modalitäten spielen. Die bisherige Forschung hat sich stark auf die Verbesserung des automatisierten Kundensupports konzentriert [4, 28]. In dieser Arbeit wird ein zielgerichtetes Dialogsystem für die faktenbasierte Informationssuche vorgestellt. Ziel dieser Arbeit ist es, geeignete Modalitäten für die Suche nach Informationen, ähnlich einer Google-Suche, zu finden. Im Gegensatz zu bisherigen Studien und als empfohlene Erweiterung von Rzepka et al. [28], werden Text und Sprache nicht als symmetrische Ein- und Ausgabemodalität gegenübergestellt, sondern auch kombiniert. Dadurch entstehen neue asymmetrische Modalitätskombinationen, wie die Spracheingabe mit anschließender Textausgabe und umgekehrt.

Die vorliegende Arbeit ist wie folgt gegliedert. In Kapitel 2 werden die bisherigen Erkenntnisse der Interaktionsmodalitätsforschung zusammengefasst. In Kapitel 3 werden eine Benutzerstudie und die dafür erstellte Anwendung vorgestellt. Es werden Hypothesen formuliert, Ergebnisse präsentiert und diskutiert. Kapitel 4 schließt die Arbeit mit einer Zusammenfassung ab.

*e-mail: leon.korkmaz@studium.uni-hamburg.de

2 Verwandte Arbeiten

Diese Arbeit schließt sich an aktuelle Studien an, die einen Vergleich von Interaktionsmodalitäten durchführen [19, 25, 28]. In diesem Kapitel werden die bisherigen Ergebnisse der Forschung zu Interaktionsmodalitäten zusammengefasst und eine theoretische Grundlage gebildet.

2.1 Task-Technology Fit

Die Task-Technology Fit (TTF)-Theorie [9] besagt, dass die Anforderungen der Aufgabe, die Eigenschaften des Nutzenden und die Fähigkeiten der Technologie übereinstimmen müssen, um positive Leistungsergebnisse zu erzielen. Wird eine hohe Übereinstimmung zwischen der Aufgabe und der Technologie wahrgenommen, wird die Technologie als effizienter und effektiver empfunden. Treten hingegen Diskrepanzen zwischen der Aufgabe, den Nutzenden oder der Technologie auf, wird die Technologie schlechter bewertet [9]. Die TTF-Theorie wurde bereits mehrfach in ähnlichen Studien [19, 25, 28] als theoretische Grundlage herangezogen und ihre Anwendbarkeit auf verschiedene Interaktionsmodalitäten [28] sowie auf die Interaktion zwischen Mensch und künstlicher Intelligenz [25] erweitert.

2.2 Interaktionsmodalitäten

Die Interaktion mit einem Dialogsystem kann in Form von Text oder Sprache stattfinden. Beide Modalitäten weisen ihre eigenen Qualitäten auf [19]. Es folgt ein Vergleich zwischen den Interaktionsmodalitäten nach verschiedenen Kriterien, die sich aus dem Einsatz der Modalitäten und den bisherigen Ergebnissen verwandter Arbeiten ergeben.

2.2.1 Anwendungsbereich und Voraussetzungen

Dialogsysteme, die auf Texteingaben reagieren, benötigen eine physische oder virtuelle Tastatur zur Eingabe von Text. Ihr Anwendungsbereich ist daher häufig auf Computer und Smartphones beschränkt. Sprache bietet sich als natürliches Interaktionsmedium an, das keinen physischen Kontakt zu einem Gerät erfordert. Dies erweitert den Kreis der verwendbaren Dialogsysteme auf freihändig bedienbare Geräte [24], wie beispielsweise Auto-Entertainment-Systeme und Heimlautsprecher. Die Interaktion mit Sprache stellt jedoch Anforderungen an die Umgebung und erfordert zusätzliche Technologien wie die Spracherkennung und Sprachsynthese. Eine laute Umgebung kann dazu führen, dass die Spracherkennung nicht richtig funktioniert [29] und die Sprachausgabe schlecht verstanden wird. Selbst unter optimalen Bedingungen stellt die Spracherkennung neben der Sprachverarbeitung eine weitere mögliche Problemquelle für das Dialogsystem dar. Le Bigot et al. [4] führten 2007 eine ähnliche Studie durch, bei der sie sprach- und textbasierte Dialogsysteme verglichen und zu dem Ergebnis kamen, dass die Interaktion über Sprache häufiger zu Fehlern im Dialog führt. Es wird darauf hingewiesen, dass diese Fehler zum Teil auf die Spracherkennung zurückzuführen sind. Durch die stetige Verbesserung der Spracherkennung [13] sollte diese Problemquelle heute einen deutlich geringeren Einfluss haben. Die Ergebnisse aus früheren Studien sind daher nur bedingt vergleichbar [30].

2.2.2 Benutzerkontrolle

Benutzerkontrolle und Freiheit ist eine der zehn Usability-Heuristiken von Jakob Nielsen [22], die als Prinzipien für eine benutzerfreundliche Gestaltung von Systemen verwendet werden. Im Vergleich zur Spracheingabe bietet die Texteingabe ein hohes Maß an Kontrolle. Die Eingabe kann im eigenen Tempo erfolgen und der Text vor dem Absenden kontrolliert und korrigiert werden [16]. Da Spracherkennungssoftware die Sprache kontinuierlich aufzeichnet und häufig die Erkennung nach Sprechpausen automatisch unterbricht, ist die Kontrolle über die Spracheingabe reduziert. Luria et al. [20] verglichen 2017 verschiedene Schnittstellen für Smart Home Geräte, darunter eine Sprachsteuerung, ein an der Wand montierter Touchscreen und ein mobiles Gerät. Die Versuchspersonen erwähnten die fehlende Kontrolle über die Sprachsteuerung und ein damit verbundenes Gefühl von Unbehagen. Auch bei der Ausgabe bietet Text mehr Kontrolle und Freiheit. Während das Hören in der Regel nur einmal möglich ist und daher ein schnelles Memorieren der gehörten Informationen erfordert, besteht geschriebener Text persistent [4]. Dieser kann im eigenen Tempo gelesen werden, erlaubt das Überspringen von Passagen und die Anwendung eigener Lesestrategien [27]. Darüber hinaus kann die Nutzung sprachbasierter Interaktion in öffentlichen Umgebungen zu einer Verletzung der Privatsphäre führen [21]. Personen fühlen sich unwohl, wenn ihre Namen oder vertrauliche Informationen, wie finanzielle oder medizinische Daten, laut ausgesprochen werden [29].

2.2.3 Natürlichkeit

Natürlicher empfundene Interaktionen mit einem Dialogsystem werden als authentischer wahrgenommen und können eine Reihe positiver Effekte hervorrufen. So ist die wahrgenommene Authentizität entscheidend darüber, ob das System als wertvoll und zufriedenstellend empfunden wird und Vertrauen aufgebaut werden kann [34]. Im direkten Vergleich hat Sprache den Vorteil, auch im zwischenmenschlichen Dialog als natürliches Kommunikationsmittel zu dienen. Dies stellt allerdings hohe Anforderungen an die Spracherkennung und Sprachsynthese, da Erkennungsfehler oder eine sehr synthetisch klingende Sprachausgabe diesen Effekt verringern würden [28]. Die Interaktion mit Text kann durch entsprechende Präsentation und Anpassung des Gesprächsstils an die Nutzenden [2] ebenfalls natürlich wirken und den Eindruck erwecken, mit einem Menschen zu schreiben. Sowohl für sprachbasierte als auch für textbasierte Dialogsysteme stehen Möglichkeiten zur Verfügung, um die Interaktion anthropomorpher zu gestalten. Textbasierte Dialogsysteme können durch eine dynamische Antwortverzögerung menschliches Verhalten wie Nachdenken und Tippen simulieren [8]. Um sprachbasierte Dialogsysteme noch natürlicher und anthropomorpher zu gestalten, kann auf die Forschung im Bereich der Psycholinguistik und Dialogforschung zurückgegriffen werden [6]. King et al. [15] konnten zeigen, dass anthropomorphe virtuelle Dialogpartner als intelligenter eingeschätzt werden: So wird anthropomorphen computergenerierten Stimmen mehr Kompetenz zugeschrieben als roboterartigen Stimmen [6]. Diese Eigenschaften können sich positiv auf die Usability von Sprachassistenten auswirken [7].

2.2.4 Kognitive Belastung

Eine theoretische Grundlage zur Vorhersage der kognitiven Belastung bildet die Media Naturalness Proposition [17]: Demnach führt ein höherer Grad an Natürlichkeit des Interaktionsmediums zu einer geringeren kognitiven Belastung [17]. Im Vergleich zu textbasierten Interaktionen werden sprachbasierte Interaktionen als natürlicher und intuitiver wahrgenommen [28]. Verschiedene Studien zeigen jedoch, dass dies auch von den Charakteristiken der Aufgabe und des Nutzenden abhängig ist [19, 25]. Rzepka et al. [28] verglichen im Jahr 2022 einen Sprachassistenten mit einem Chatbot in einer zielgerichteten (Suche nach einem Restaurant mit bestimmten Kriterien) und einer nicht-zielgerichteten explorativen Suchaufgabe (Suche nach allen Restaurants ohne bestimmte Kriterien). Die Ergebnisse zeigen, dass die sprachbasierte Interaktion zu einer geringeren kognitiven Belastung beiträgt und sich dieser Effekt nicht zwischen zielgerichteten und explorativen Aufgaben unterscheidet. Die zielgerichtete Aufgabe reduzierte jedoch die Belastung bei der textbasierten Interaktion. Le Bigot et al. [4] stellten eine höhere kognitive Belastung durch die sprachbasierte Interaktion fest, insbesondere bei komplexeren Aufgaben. Als Ursache wird die höhere Belastung des Arbeitsgedächtnisses genannt, da die Ausgabe von Sprache, wie in 2.2.2 beschrieben, nur temporär zur Verfügung steht und so ein Memorieren der gehörten Informationen notwendig ist.

2.2.5 Effizienz

Die Effizienz einer Modalität kann subjektiv durch die wahrgenommene Effizienz und objektiv über Messparameter wie die Interaktionszeit und die Fehlerrate gemessen werden. Ruan et al. [26] verglichen 2017 die Eingabezeit kurzer Texte über eine Spracherkennung und eine virtuelle Tastatur. Mit der Spracherkennung konnte eine um den Faktor 2.93 schnellere Eingabegeschwindigkeit für englische Texte erzielt werden. Le Bigot et al. [4] erhoben in ihrer Studie neben den Bearbeitungszeiten auch die Fehlerraten. Die Ergebnisse zeigen eine kürzere Bearbeitungszeit und damit eine höhere Produktivität bei der Verwendung der sprachbasierten Eingabe, aber auch eine viermal höhere Fehlerrate im Vergleich zur rein textbasierten Interaktion. Le Bigot et al. schlussfolgern aus dieser Fehlerrate, dass Sprache im Vergleich zu Text eine geringe Effizienz aufweist, wobei die Kombination von Spracheingabe und Textausgabe die besten Bearbeitungszeiten bei moderater Fehlerrate erzielt. Rzepka et al. [28] erhoben die Effizienz anhand von Fragebögen und kamen zu dem Ergebnis, dass die Interaktion mit Sprache als effizienter wahrgenommen wird.

2.2.6 Freude

Pal et al. [24] verglichen im Jahr 2020 verschiedene Modelle und Faktoren zur Erklärung von Technologieakzeptanz. Dabei zeigte sich, dass Freude im Vergleich zu Nützlichkeit und subjektiven Normen den größten Einfluss auf die Akzeptanz sprachbasierter Technologien hat [24]. Rzepka et al. [28] kamen zu dem Ergebnis, dass die Interaktion mit Sprache zu mehr Freude führt, als die Interaktion mit Text. Dabei spielte es keine Rolle, ob es sich um eine zielgerichtete oder explorative Aufgabe handelte. Als Erklärungsansatz wird der *Novelty Effect* herangezogen, da es sich bei der Interaktion mit Sprache um eine relativ neue Interaktionsmodalität handelt [28].

2.2.7 Zusammenfassung

Welche Modalität in einem bestimmten Kontext verwendet werden sollte, hängt von verschiedenen Faktoren ab. Erfolgt die Bedienung eines Dialogsystems freihändig, ist die Verwendung von Sprache notwendig, erfolgt die Interaktion mit dem System in einer lauten Umgebung, sollte auf Text zurückgegriffen werden. Insbesondere in nicht zielgerichteten Interaktionen kann Sprache als natürliche und intuitive Modalität die kognitive Belastung reduzieren [17] und einen höheren Grad an Natürlichkeit und Freude [28] mit sich bringen. In zielgerichteten Interaktionen kann auch Text durch seine Kontrollierbarkeit die kognitive Belastung reduzieren und ist weniger anfällig für Sprachverarbeitungsfehler. Neben der Betrachtung des Kontextes ist auch eine separate Betrachtung der Ein- und Ausgabemodalität sinnvoll. Sprache hat sich in den genannten Studien als effiziente Eingabemodalität erwiesen [4, 26, 28]. Text hat sich aufgrund seiner hohen Benutzerkontrolle als effiziente Ausgabemodalität erwiesen. So führte die Kombination von Spracheingabe und Textausgabe in Le Bigots Studie zu den besten Bearbeitungszeiten [4].

In der vorgestellten Literatur wurden asymmetrischen Modalitätskombinationen wenig Aufmerksamkeit geschenkt. Daher wird in dieser Arbeit ein Schwerpunkt auf die Kombination unterschiedlicher Modalitäten gelegt. Insbesondere die Kombination von Spracheingabe und Textausgabe könnte eine effiziente und kognitiv wenig belastende Interaktionsmöglichkeit bieten und damit einen hohen Task-Technology Fit [9] erzielen. Von großem Interesse sind auch die Auswirkungen auf Usability und User Experience solcher asymmetrischen Kombinationen, die in der bisherigen Literatur noch nicht diskutiert wurden.

3 Benutzerstudie

In diesem Kapitel wird die Benutzerstudie beschrieben, in der die verschiedenen Ein- und Ausgabemodalitäten verglichen wurden.

3.1 Versuchspersonen

28 Personen (19 weiblich, 8 männlich, 1 Nicht-binär; Alter: 68% 18-24, 28% 25-34, 4% 55-64) haben an der Studie teilgenommen. Alle Teilnehmenden wurden über das Versuchspersonenstunden-System SONA der Universität Hamburg rekrutiert. Um Sprachbarrieren bei der Spracherkennung zu vermeiden, wurde als Kriterium Deutsch auf muttersprachlichem Niveau (C2) fließend in Wort und Schrift vorausgesetzt. 22 Personen gaben an, Vorerfahrungen mit Sprachassistenzsystemen (Alexa, Siri, Google) zu haben. Davon 5 mindestens einmal täglich, 4 wöchentlich, 4 monatlich, 9 jährlich. 25 Personen gaben an, Vorerfahrungen mit Chatassistenzsystemen (Kundensupport, KI-Tools) zu haben. Davon 4 mindestens einmal täglich, 10 wöchentlich, 6 monatlich, 5 jährlich.

3.2 Material

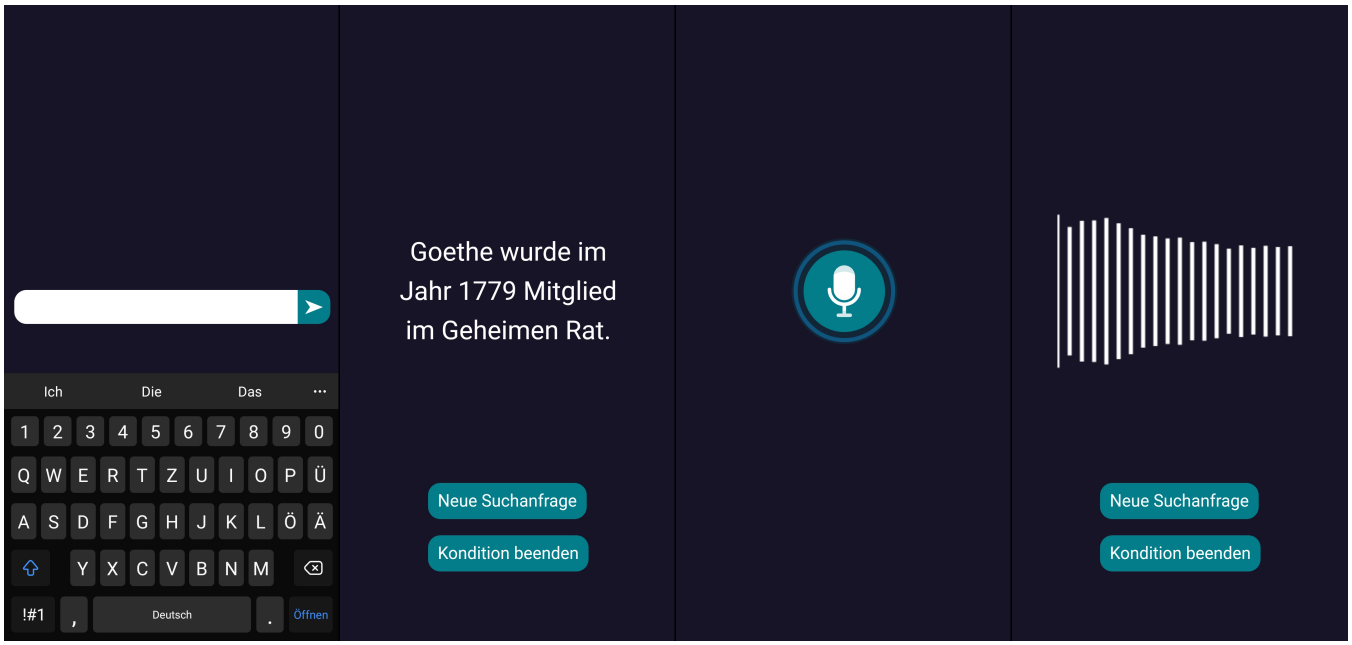
Um die Kombinationen aus Spracheingabe und Textausgabe (S-T), Texteingabe und Sprachausgabe (T-S), sowie die rein textbasierten (T-T) und rein sprachbasierten (S-S) Kombinationen miteinander vergleichen zu können, wurde ein zielgerichtetes Dialogsystem auf einem Smartphone erstellt, mit dem die Versuchspersonen vorgefertigte Fragen beantworten mussten.

3.2.1 Anwendung

Die Anwendung wurde als Web-Applikation mit dem Javascript-Framework SvelteKit erstellt und für mobile Endgeräte optimiert. Für die Verarbeitung der Eingaben wurde ein Google Dialogflow Agent verwendet. Google Dialogflow ist eine Plattform, die KI-basierte Konversationsagenten bereitstellt, mit denen Chatbots und andere Dialogsysteme realisiert werden können. Dialogflow wurde bereits in ähnlichen Studien [28] eingesetzt. Der Agent wurde zuvor mit verschiedenen Variationen der Fragen trainiert und gab einen gleichbleibenden Antwortsatz zurück. Da Google Dialogflow nur einen bestimmten Audiocodec akzeptiert, der vom verwendeten Mikrofon nicht unterstützt wurde, erfolgte keine direkte Übertragung der Sprachdaten. Stattdessen wurde die Spracherkennung über die im Chrome-Browser implementierte Google Webspeech API realisiert und das Ergebnis als Text an den Dialogflow-Agenten übermittelt. Der Agent führt ein Intent Matching durch und sendet die Antwort als Text und Audiodatei zurück. Als Stimme wurde die männliche deutsche Variante von Neural2 verwendet. Diese stellt die neueste Entwicklung der Sprachsynthese von Google dar und kann als Weiterentwicklung des bereits als sehr natürlich empfundenen Google WaveNet Modells eingeordnet werden [32].

Die Anwendung wurde so konzipiert, dass sie die Versuchsperson durch die Studie begleitet. Ein Durchlauf der Anwendung besteht jeweils aus den vier genannten Modalitätskombinationen, nachfolgend Konditionen genannt. Eine Kondition kann erst abgeschlossen werden, sobald mindestens drei Suchanfragen in der Anwendung gestellt wurden. Zu Beginn einer Kondition wird ein 10 sekündiger Countdown gestartet, der auf die nachfolgende Eingabemodalität hinweist: *Bitte gebe deine Frage nachfolgend in das Textfeld ein* oder *Bitte spreche deine Frage nachfolgend in das Mikrofon*. Nach der Ausgabe besteht die Möglichkeit, eine neue Suche in der aktuellen Kondition zu tätigen oder die Kondition zu beenden. Wird die Kondition beendet, erscheint der Hinweis: *Fülle nun den Fragebogen aus und drücke danach auf weiter*. Ein Klick auf *weiter* startet eine neue Kondition, bis alle vier Konditionen abgeschlossen sind. Die Anwendung erfasst die Eingabezeit nach Ablauf des Countdowns oder nach dem Klick auf eine *Neue Suchanfrage*, bis die Anfrage abgeschickt oder die Spracherkennung beendet wird. Ebenfalls erfasst wird die Ausgabezeit ab dem Zeitpunkt der Präsentation der Antwort, bis eine weitere Suchanfrage gestellt oder die Kondition beendet wird. Eine Erläuterung des Ablaufs folgt in 3.3 Methodik.

Abbildung 1: Darstellung der Ein- und Ausgabemodi in der mobilen Anwendung



a) Texteingabe b) Textausgabe c) Spracheingabe d) Sprachausgabe

Für die verschiedenen Modalitäten ergeben sich zwei Eingabemodi und zwei Ausgabemodi. Diese werden von der Anwendung visuell wie folgt dargestellt (Abbildung 1):

a) *Texteingabe.* Die Texteingabe findet über ein Textfeld statt. Mithilfe der Öffnen-Taste oder einem Klick auf den mit einem Pfeil gekennzeichneten Submit-Button wird der Text an den Agenten übermittelt.

b) *Spracheingabe.* Die Spracheingabe findet über ein Mikrofon statt. Visuell wird dieses durch ein animiertes Mikrofonsymbol dargestellt. Sobald die Spracherkennung aktiv ist, ertönt ein Signalton. Erfolgt eine Sprechpause, wird die Spracherkennung beendet und ein weiterer Signalton ertönt.

c) *Textausgabe.* Die Textausgabe erfolgt in Schriftform zentriert auf dem Bildschirm.

d) *Sprachausgabe.* Die Sprachausgabe erfolgt über einen Lautsprecher und wird visuell durch eine Wellenform dargestellt.

3.2.2 Hardware

Als Smartphone kam ein Samsung Galaxy A51 zum Einsatz. Da die Anwendung im Vollbildmodus gestartet wurde, belegte sie das gesamte 6.5 Zoll große Touch Display. Texteingaben erfolgten über die virtuelle Samsung-Tastatur. Die eingeschaltete Autokorrektur der Tastatur wurde nach jedem Durchlauf zurückgesetzt, um gelernte Vorschläge zu vermeiden. Die Spracheingabe erfolgte über das integrierte Mikrofon des Smartphones. Die Sprachausgabe erfolgte über die integrierten Lautsprecher bei 80% der maximalen Lautstärke.

3.2.3 Frageblöcke

Für die Anwendung wurden vier Frageblöcke zum Thema Goethe erstellt. In jedem Frageblock wurde nach dem Namen einer Person aus Goethes Leben, dem Jahr eines Ereignisses und einer Stadt, in der ein Ereignis stattfand, gefragt. Die Fragen wurden mithilfe von ChatGPT generiert und sollten möglichst kurze und wenig bekannte Fakten aus Goethes Leben abfragen. Um den Versuchspersonen keine direkte Frage vorzugeben, wurden diese als neutral zu erfragende Fakten formuliert. Die vier Frageblöcke wurden auf DIN A4 Blätter gedruckt und in kleine Aufgabenzettel mit jeweils einem Frageblock aufgeteilt (Abbildung 2). Durch das analoge Aufschreiben der Antworten, musste keine weitere Anwendung auf einem technischen Gerät geöffnet werden, die von der mobilen Anwendung hätte ablenken können.

Abbildung 2: Beispiel Aufgabenzettel zum Thema Goethe

Jahr der Ernennung Goethes zum Geheimen Rat: _____

Name Goethes erstgeborener Sohn: _____

Stadt, die Goethe in Italien besonders faszinierte: _____

3.3 Methodik

Für die Studie wurde ein 2 (Eingabemodalität: Text vs. Sprache) x 2 (Ausgabemodalität: Text vs. Sprache) Within-Subject-Design gewählt. Die Versuchspersonen durchliefen alle Konditionen in randomisierter Reihenfolge. Diese ergab sich aus der Permutation der vier Konditionen (24) und dem lateinischen Quadrat (4).

3.3.1 Durchführung der Studie

Die Studie wurde in einzelnen Laborsitzungen durchgeführt, um kontrollierte Bedingungen zu schaffen und mögliche Hemmungen bei der sprachbasierten Interaktion mit dem Gerät zu verringern. Nach dem Ausfüllen der Einwilligungserklärung wurde eine kurze Einleitung zum generellen Ablauf der Studie gegeben und das Smartphone mit der Anwendung präsentiert.

Vor der Durchführung der Studie wurde ein Testdurchlauf mit vier Fragen gestartet, um die Versuchspersonen mit der Anwendung vertraut zu machen. Während dieses Testdurchlaufs folgte eine Instruktion, in der die Versuchsperson darüber aufgeklärt wurde, die geforderte Information in einer eigenen freien Formulierung an das System zu stellen. Es wurde darauf hingewiesen, dass nur die geforderten Informationen (Jahreszahl, Name einer Person, Name einer Stadt) im folgenden Durchlauf notiert werden müssten und dass im Falle eines Fehlversuchs (unpassende oder keine Antwort) die Frage umformuliert und erneut abgeschickt werden sollte.

Es folgte die Aushändigung des ersten Aufgabenzettels. Die Versuchspersonen wurden gebeten, sich die Fragen anzusehen und, wenn sie sich dazu bereit fühlten, dem in 3.2.1 beschriebenen Ablauf der Anwendung zu folgen. Nach Beendigung einer Kondition, wurde der Versuchsperson ein Laptop zur Beantwortung der Fragebögen überreicht. Anschließend wurde ein neuer Aufgabenzettel mit den nächsten drei Fragen ausgehändigt und in die nächste Kondition gestartet. Abschließend wurden in einem Fragebogen demographische Daten, Vorerfahrungen mit Assistenzsystemen, sowie eine Rangfolge der Konditionen nach persönlichen Präferenzen mit Begründung erhoben.

3.3.2 Messwerte und Hypothesen

Nach jedem Durchlauf einer Kondition wurden die Versuchspersonen gebeten, eine Reihe von Fragebögen auszufüllen. Neben der vom Programm erfassten Bearbeitungszeit konnten so auch die empfundene Arbeitsbelastung, die Usability, die User Experience und die Natürlichkeit des Antwortverhaltens erfasst werden. Basierend auf den Ergebnissen verwandter Arbeiten (Kapitel 2) haben sich Spracheingabe und Textausgabe in verschiedenen Kriterien als vorteilhaft erwiesen. Eine asymmetrische Modalitätskombination aus Spracheingabe und Textausgabe könnte sich somit als effektive Alternative zu symmetrischen Kombinationen behaupten. Um diese Annahme zu testen, wurde eine Reihe von Hypothesen aufgestellt.

Die erste Hypothese H1 knüpft an die Erkenntnisse von Le Bigot [4] zur Erfassung der Effizienz an und überprüft, ob das Ergebnis auch auf das vorgestellte zielgerichtete System übertragbar ist. Als objektiver Messparameter dient die Bearbeitungszeit jeder Frage, die durch die Anwendung erfasst wird.

H1a: Die Verwendung von Sprache als Eingabemodalität führt zu einer höheren Effizienz.
H1b: Die Verwendung von Text als Ausgabemodalität führt zu einer höheren Effizienz.
H1c: Die Kombination von Spracheingabe und Textausgabe führt im Vergleich zu anderen Kombinationen zu der höchsten Effizienz

Die zweite Hypothese H2 knüpft ebenfalls an die bisherigen Ergebnisse [4] an. Zur Erfassung der kognitiven Belastung wurde der NASA-TLX Fragebogen [12] herangezogen. Dieser wurde bereits von Le Bigot [4] zur Erfassung der kognitiven Belastung eingesetzt. Der NASA-TLX kann die Arbeitsbelastung in sechs Dimensionen erfassen. Dabei werden Werte von 0 bis 100 ermittelt, wobei höhere Werte einer höheren Belastung entsprechen. In dieser Studie wurde eine deutsche Replikation mit einer 21-Punkte-Skala verwendet.

H2a: Die Verwendung von Sprache als Eingabemodalität führt zu einer geringeren kognitiven Belastung.
H2b: Die Verwendung von Text als Ausgabemodalität führt zu einer geringeren kognitiven Belastung.
H2c: Die Kombination von Spracheingabe und Textausgabe führt im Vergleich zu anderen Kombinationen zu der geringsten kognitiven Belastung.

Die vermuteten Vorteile von Spracheingabe und Textausgabe könnten sich auch positiv auf die User Experience und Usability auswirken. Zudem könnten asymmetrische Kombinationen als neuartig und origineller empfunden werden und dadurch eine höhere Präferenz erzielen. Es ist aber auch denkbar, dass die Asymmetrie als Inkonsistenz im System wahrgenommen wird.

Die dritte Hypothese H3 bezieht sich auf die wahrgenommene Usability der Kombinationen. Die Usability wurde mit dem System Usability Scale (SUS) [5] Fragebogen erhoben. Dieser ermittelt eine Punktzahl von 0 bis 100, wobei eine höhere Punktzahl für eine höhere Usability steht.

H3: Die Kombination aus Spracheingabe und Textausgabe erzielt im Vergleich zu anderen Kombinationen eine höhere Usability.

Die vierte Hypothese H4 bezieht sich auf die User Experience. Die Erfassung erfolgte mit der Kurzversion des User Experience Questionnaire (UEQ-S) [31]. Dieser unterscheidet in pragmatische und hedonische Qualität. Die pragmatische Qualität setzt sich daraus zusammen, wie übersichtlich, effizient, einfach und unterstützend das System empfunden wird. Die hedonische Qualität setzt sich daraus zusammen wie interessant, spannend, originell und neuartig das System empfunden wird [31]. Der Wertebereich des UEQ-S reicht von -3 bis 3. Ein höherer Wert steht für eine höhere Qualität.

H4a: Die Kombination aus Spracheingabe und Textausgabe erzielt im Vergleich zu allen anderen Kombinationen eine höhere pragmatische Qualität.
H4b: Die Kombination aus Spracheingabe und Textausgabe erzielt im Vergleich zu allen anderen Kombinationen eine höhere hedonische Qualität.

Die fünfte Hypothese H5 bezieht sich auf die wahrgenommene Natürlichkeit des Antwortverhaltens. Diese beschränkt sich auf die Ausgabemodalität. Zur Beurteilung der Natürlichkeit des Antwortverhaltens wurde die erste Kategorie Anthropomorphismus der Godspeed Questionnaire Series (GQS) [3] herangezogen. Der Wertebereich liegt zwischen 0 bis 5. Je höher der Wert, desto anthropomorpher das wird das Verhalten wahrgenommen.

H5: Die Kombinationen mit Sprachausgabe werden als natürlicher empfunden als die Kombinationen mit Textausgabe.

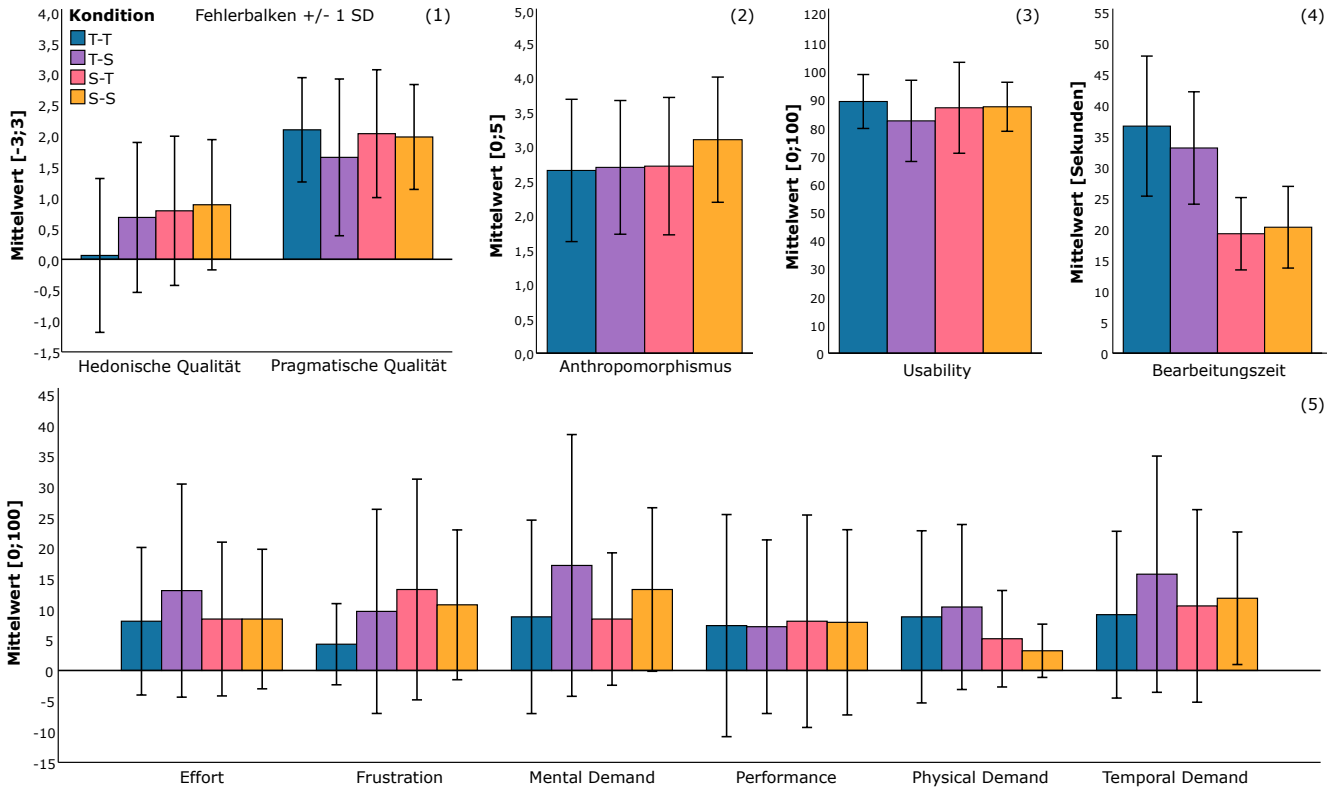
3.4 Ergebnisse

Die vorgestellten Hypothesen wurden mithilfe verschiedener statistischer Analysen auf ihre Gültigkeit überprüft. Die Daten wurden vorab bereinigt und auf Normalverteilung geprüft. Vereinzelt traten verzerrte Audioausgaben oder ein fehlerhaftes Intent Matching des Dialogflow-Agenten auf. Die fehlerhaften Durchläufe wurden aus dem Datensatz entfernt. Da eine Kondition nur mit drei korrekten Durchläufen abgeschlossen werden konnte, hat die Bereinigung keine Auswirkungen auf den Datensatz.

3.4.1 Effizienz

Bearbeitungszeit. Die Anwendung erfasste die Bearbeitungszeit jedes Durchlaufs in Millisekunden. Zur Auswertung der Messwerte wurde eine zweifaktorielle ANOVA mit Messwiederholung durchgeführt. Um die Voraussetzung der Normalverteilung zu erfüllen, wurden die Daten zuvor mit $x_t = log(x)$ transformiert. Die ANOVA ermittelte einen Haupteffekt auf der Eingabemodalität ($F(1,83) = 363.477, p =< 0.001, \eta_p^2 = 0.814$), keinen Haupteffekt auf der Ausgabemodalität ($F(1,83) = 1.001, p = 0.320, \eta_p^2 = 0.012$) aber einen Interaktionseffekt $F(1,83) = 5.053, p = 0.027, \eta_p^2 = 0.057$).

Abbildung 3: Grafische Darstellung der Ergebnisse aus dem UEQ-S (1), GQS (2), SUS (3), NASA-TLX (5) und den Bearbeitungszeiten (4)



Da der Interaktionseffekt die Aussagekraft der Haupteffekte einschränkt, wurde eine Analyse der einfachen Haupteffekte durchgeführt. Diese zeigt, dass die Bearbeitungszeit bei Verwendung der Spracheingabe signifikant kürzer ist, unabhängig davon, ob diese mit der Textausgabe ($p =< 0.001$) oder der Sprachausgabe ($p =< 0.001$) kombiniert wird. Im Vergleich zur Sprachausgabe führt die Textausgabe in Kombination mit der Texteingabe zu einer signifikant längeren Bearbeitungszeit ($p = 0.030$), mit der Spracheingabe gibt es keinen signifikanten Unterschied ($p = 0.312$). Um den Interaktionseffekt im Bezug auf die Hypothese H1c genauer zu untersuchen, wurde eine einfaktorielle ANOVA auf dem Faktor Kondition mit Bonferroni-Korrektur durchgeführt. Die Kondition S-T konnte sich signifikant von T-T ($p =< 0.001$) und T-S ($p =< 0.001$) unterscheiden. Zu S-S ($p = 1.000$) wurde kein signifikanter Unterschied festgestellt.

Wahrgenommene Effizienz. Zur Bestimmung der wahrgenommenen Effizienz wurden die Dimensionen *Effort* und *Performance* des NASA-TLX Fragebogens [12] verwendet. Da die Daten nicht normalverteilt waren, erfolgte die Auswertung mittels eines nicht-parametrischen Friedman-Tests. Sowohl für *Effort* ($\chi^2(3) = 3.703, p = 0.299$), als auch für *Performance* ($\chi^2(3) = 0.775, p = 0.868$) konnten keine statistisch signifikanten Unterschiede gefunden werden.

3.4.2 Kognitive Belastung

Zur Ermittlung der kognitiven Belastung wurde die Dimension *Mental Demand* des NASA-TLX Fragebogens verwendet. Die Auswertung erfolgte mittels eines nicht-parametrischen Friedman-Tests. Es konnte ein statistisch signifikanter Unterschied in der wahrgenommenen mentalen Beanspruchung festgestellt werden ($\chi^2(3) = 8.644, p = 0.032$). Die Post-hoc-Analyse mit Wilcoxon Signed-Rank-Tests wurde mit einer Bonferroni-Korrektur durchgeführt und ergab ein Signifikanzniveau von $p < 0.0083$. Mit dem neuen Signifikanzniveau konnte kein signifikanter Unterschied mehr festgestellt werden. Da sich Frustration und Zeitdruck auch auf die kognitive Belastung auswirken können [11], wurde die gleiche Auswertung auf die Dimensionen *Frustration* und *Temporal Demand* angewendet. Es konnte ein statistisch signifikanter Unterschied in der wahrgenommenen Frustration festgestellt werden ($\chi^2(3) = 9.937, p = 0.019$). Die Post-hoc-Analyse ergab jedoch aufgrund des neuen Signifikanzniveaus von $p < 0.0083$ nach Bonferroni-Korrektur keinen signifikanten Unterschied zwischen den Gruppen. Für den wahrgenommenen Zeitdruck konnte ebenfalls kein signifikanter Unterschied festgestellt werden ($\chi^2(3) = 3.754, p = 0.290$).

3.4.3 Usability

Die Punktzahl aus dem SUS Fragebogen [5] wurde mittels zweifaktorieller ANOVA mit Messwiederholung ausgewertet. Um die Annahme der Normalverteilung annähernd zu erfüllen, wurden die Werte auf 0-1 normiert und eine $x_t = arcsin(x)^{(\frac{1}{2})}$-Transformation durchgeführt. Ein Blick auf das Histogramm und das QQ-Diagramm zeigte eine leichte negative Schiefe (0.508) und Kurtosis (0.865). Diese Abweichungen sollten akzeptabel sein, da die ANOVA robust gegenüber leichten Abweichungen von der Normalverteilung ist [23]. Die Analyse ergab einen statistisch signifikanten Haupteffekt auf der Ausgabemodalität ($F(1,27) = 5.950, p = 0.022, \eta_p^2 = 0.181$). Es wurden keine weiteren signifikanten Effekte gefunden.

3.4.4 User Experience

Aus dem UEQ-S Fragebogen [31] wurden Werte für die pragmatische und für die hedonische Qualität ermittelt. Die Werte wurden mittels zweifaktorieller ANOVA mit Messwiederholung ausgewertet. Um die Werte der pragmatischen Qualität annähernd normalverteilt zu erhalten, wurden diese auf 0-1 normiert und eine $x_t = arcsin(x)^{(\frac{1}{2})}$-Transformation durchgeführt. Auch hier zeigten

das Histogramm und das QQ-Diagramm eine tolerierbare negative Schiefe (0.590) und Kurtosis (0.289). Für die pragmatische Qualität konnte kein statistisch signifikanter Effekt festgestellt werden. Die Analyse der hedonischen Qualität ergab einen signifikanten Haupteffekt auf der Eingabemodalität ($F(1,27) = 11.775, p = 0.002, \eta_p^2 = 0.304$), einen signifikanten Haupteffekt auf der Ausgabemodalität ($F(1,27) = 8.773, p = 0.006, \eta_p^2 = 0.245$) und einen signifikanten Interaktionseffekt ($F(1,27) = 6.704, p = 0.015, \eta_p^2 = 0.199$). Die Folgeanalyse der einfachen Haupteffekte zeigt, dass die Spracheingabe unter Verwendung der Textausgabe zu einer signifikant höheren hedonischen Qualität führt als die Texteingabe ($p = 0.024$), nicht aber unter Verwendung der Sprachausgabe ($p = 0.518$). Im Vergleich zur Textausgabe kann sich die Sprachausgabe weder in Kombination mit der Texteingabe ($p = 0.054$) noch mit der Spracheingabe ($p = 0.757$) signifikant unterscheiden. Um den Interaktionseffekt im Bezug auf die Hypothese H4b genauer zu untersuchen, wurde eine einfaktorielle ANOVA auf dem Faktor Kondition mit Bonferroni-Korrektur durchgeführt. Die Kondition S-T unterschied sich signifikant von T-T ($p =< 0.001$), jedoch nicht von T-S ($p = 1.000$) oder S-S ($p = 1.000$).

3.4.5 Natürlichkeit

Die Werte aus der Kategorie Anthropomorphismus der Godspeed Questionnaire Series [3] wurden mittels zweifaktorieller ANOVA mit Messwiederholung ausgewertet. Die Analyse ergab einen signifikanten Haupteffekt auf der Eingabemodalität ($F(1,27) = 10.890, p = 0.003, \eta_p^2 = 0.287$), keinen Haupteffekt auf der Ausgabemodalität ($F(1,27) = 3.358, p = 0.078, \eta_p^2 = 0.111$) und keinen Interaktionseffekt ($F(1,27) = 2.676, p = 0.114, \eta_p^2 = 0.090$).

3.4.6 Präferenz

Abschließend konnten die Teilnehmenden die verschiedenen Konditionen nach ihrer persönlichen Präferenz in eine Rangfolge bringen. Dabei bevorzugten 39.3% T-T, 35.7% S-T, 21.4% S-S und 3.6% T-S. Um zu überprüfen, ob verschiedene Szenarien außerhalb der Laborbedingungen einen Einfluss auf die Präferenzen haben, wurden zusätzliche Szenarien präsentiert. Die Ergebnisse sind in Tabelle 1 zusammengefasst.

1. *Szenario Zuhause: Stellen Sie sich vor, Sie würden Zuhause an ihrem Smartphone sitzen und möchten Informationen über ein Museum herausfinden, welches Sie morgen besuchen möchten.*

2. *Szenario Museum: Stellen Sie sich vor, Sie würden sich in einem Goethe-Museum befinden und möchten Informationen zu Goethe herausfinden.*

Tabelle 1: Absolute und relative Anzahl an Stimmen für die bevorzugte Konditionen in den verschiedenen Szenarien.

	T - T	T - S	S - T	S - S
Labor	11 (39.3%)	1 (3.6%)	10 (35.7%)	6 (21.4%)
Zuhause	17 (60.7%)	0 (0%)	7 (25%)	4 (14.3%)
Museum	15 (53.6%)	5 (17.9%)	6 (21,4%)	2 (7.1%)
Gesamt	43 (51.2%)	6 (7.1%)	23 (27.4%)	12 (14.3%)

3.5 Diskussion

Die Ergebnisse zeigen geringere Unterschiede zwischen den Konditionen als erwartet. Im Folgenden werden die Hypothesen diskutiert, die Ergebnisse eingeordnet und Erklärungsansätze vorgestellt.

3.5.1 Effizienz

Auf der Grundlage des NASA-TLX Fragebogens konnten keine signifikanten Unterschiede in der wahrgenommenen Leistung oder Anstrengung festgestellt werden. Die ANOVA zeigte jedoch signifikante Unterschiede in den Bearbeitungszeiten. Eine Folgeanalyse bestätigte, dass die Spracheingabe zu signifikant kürzeren Bearbeitungszeiten führte, sodass die Hypothese (H1a) zumindest über objektive Messparameter angenommen werden kann. Die Konditionen mit Spracheingabe wurden durchschnittlich um den Faktor 1.76 schneller abgeschlossen (Abbildung 3). Hinsichtlich der Ausgabemodalität konnte in der Folgeanalyse ein signifikanter Unterschied festgestellt werden, der jedoch gegen die Hypothese (H1b) spricht. Dass Le Bigot [4] einen signifikanten Unterschied zugunsten der Textausgabe finden konnte, könnte auf längere und umfangreichere Antwortsätze zurückzuführen sein. Diese würden die Vorteile der Textausgabe, wie das selektive Herausfiltern von Informationen, verstärken. Die Kondition Spracheingabe und Textausgabe führte zwar zu den kürzesten Bearbeitungszeiten, kann sich aber nicht signifikant von der rein sprachlichen Kondition unterscheiden, sodass auch die Hypothese (H1c) nicht ausreichend unterstützt wird.

3.5.2 Kognitive Belastung

Keine der aufgestellten Hypothesen (H2a, H2b, H2c) konnte bestätigt werden. Es sind zwar Unterschiede zwischen den Konditionen zugunsten der Hypothesen erkennbar (Abbildung 3), diese sind jedoch nicht signifikant. Verschiedene Faktoren können dieses Ergebnis erklären. In verwandten Studien wurden sowohl positive als auch negative Effekte bei der Interaktion über Spracheingabe und Textausgabe festgestellt (siehe 2.2.4). Möglicherweise heben sich diese Effekte gegenseitig auf. So reduziert die Natürlichkeit gemäß der Media Naturalness Proposition [17] die kognitive Belastung, während die verringerte Benutzerkontrolle bei der Spracheingabe die kognitive Belastung erhöhen könnte. Denkbar wäre auch, dass ein unvorhergesehener Modalitätswechsel bei asymmetrischen Kombinationen die kognitive Belastung erhöht. Einige Versuchspersonen gaben an, den Modalitätswechsel als irreführend, irritierend und weniger intuitiv empfunden zu haben.

3.5.3 Usability

Bezüglich der Usability konnte sich keine der Konditionen als signifikant besser erweisen. Der gefundene Haupteffekt auf der Ausgabemodalität weist auf eine verbesserte Usability durch die Textausgabe hin. Weitere Effekte konnten nicht festgestellt werden. Die Hypothese (H3) wird daher nicht angenommen. Der SUS-Fragebogen enthält Fragen zur Konsistenz und zur wahrgenommenen Verwirrung durch das System. Es ist zu vermuten, dass asymmetrische Modalitätskombinationen in dieser Dimension schlechter abschneiden als symmetrische Modalitätskombinationen. Darauf deuten die Begründungen der Versuchspersonen hin (siehe 3.5.2).

3.5.4 User Experience

Die User Experience wurde in pragmatische und hedonische Qualität unterteilt. Bezüglich der pragmatischen Qualität konnten keine signifikanten Unterschiede festgestellt werden. Die Hypothese (H4a) wird somit nicht angenommen. Da sich die pragmatische Qualität an Parametern wie Effizienz und Usability orientiert, fügt sich dieses Ergebnis in die bisherigen Resultate ein. Für die hedonische Qualität konnten sowohl Haupteffekte auf der Ein- und Ausgabemodalität als auch ein Interaktionseffekt gefunden werden. Die Folgeanalyse zeigte einen signifikanten Unterschied auf der Eingabemodalität zwischen T-T und S-T, jedoch nicht auf der Ausgabemodalität. Mit Blick auf die Abbildung 3 lässt sich der Trend erkennen, dass der Einsatz von Sprache einen positiven Effekt auf die hedonische Qualität hat. Dies unterstützt die Ergebnisse von Rzepka et al. [28], die herausfanden, dass die Interaktion mit Sprache am meisten Freude bereitet.

Sprache ist im Vergleich zu Text eine relativ neue Interaktionsmodalität und kann durch moderne Sprachsynthese sehr menschlich wirken, was die Interaktion interessanter und spannender gestalten kann. Der von Rzepka et al. beschriebene *Novelity Effect* könnte dazu beitragen, dass die rein sprachliche Interaktion, sowie die neu vorgestellten asymmetrischen Modalitätskombinationen eine so hohe hedonische Qualität im Vergleich zur textbasierten Interaktion aufweisen. Entgegen der Annahme, dass dieser Effekt bei asymmetrischen Kombinationen noch stärker ausfällt, konnte die rein sprachliche Kondition die höchste hedonische Qualität aufweisen. Es konnten jedoch keine signifikanten Unterschiede zwischen den drei Kombinationen mit hoher hedonischer Qualität festgestellt werden, sodass die Daten Hypothese (H4b) nicht unterstützen.

3.5.5 Natürlichkeit

Da sich die Erfassung der Natürlichkeit auf das Antwortverhalten beschränkt, ist der Haupteffekt auf der Eingabemodalität zu vernachlässigen. Entgegen der Erwartung konnte kein Haupteffekt auf der Ausgabemodalität gefunden werden, sodass die Hypothese (H5) nicht angenommen werden kann. Da die rein sprachbasierte Kondition besser als die Texteingabe mit Sprachausgabe abgeschnitten hat (Abbildung 3), könnte dies darauf hinweisen, dass asymmetrische Modalitätskombinationen die Wahrnehmung der Natürlichkeit des Antwortverhaltens reduzieren.

3.5.6 Präferenz

Im Gegensatz zu den bisherigen Ergebnissen wird die rein textbasierte Kondition in allen vorgestellten Szenarien bevorzugt. Dies widerspricht insbesondere den Ergebnissen zur hedonischen Qualität, von der angenommen wurde, dass sie die Präferenz positiv beeinflusst. Die Spracheingabe mit Textausgabe kann nur im Labor mit 10 zu 11 Stimmen als gleichwertige Alternative zur textbasierten Kondition angesehen werden. In den weiteren vorgestellten Szenarien setzte sich die textbasierte Kondition mit über 50% der Stimmen durch. Ein Blick auf die Abbildung 3 zeigt für diese das geringste Frustrations- und Zeitdruckempfinden, sowie die höchsten Werte für Usability und Pragmatische Qualität. Die Werte unterscheiden sich jedoch nicht signifikant von den anderen Konditionen, unterstützen aber die von den Teilnehmenden angegebenen Begründungen. Die Teilnehmenden schätzten den hohen Grad an Freiheit und Kontrolle bei der textbasierten Kondition: *„Beim Tippen kann man ruhiger und länger nachdenken“*, *„Ich kann deutlich schneller lesen als dass ich zuhören kann“*. Als effektivste Modalitätskombination wurde die Spracheingabe mit Textausgabe genannt: *„Spracheingabe - Textausgabe ist am effizientesten, da ich schneller spreche als schreibe und schneller lese als höre“*. Dennoch wurde mehrfach betont, dass symmetrische Modalitätskombinationen vertrauter wirken: *„Gleiche Methoden von Input und Output fühlen sich normaler und erwarteter an, bei unterschiedlichen Methoden war ich immer kurz irritiert und habe mich unter Druck gesetzt gefühlt“*. Im Szenario Zuhause gaben einige Personen, die ihre Präferenz von Spracheingabe auf Texteingabe geändert haben, an, dass sie die Texteingabe für komplexere Eingaben bevorzugen: *„Texteingabe ist für komplexere Eingaben meiner Meinung nach besser, da man mehr Zeit zum Überlegen hat und man Fragen dadurch besser formulieren kann“*. Im Szenario Museum wurde die sprachbasierte Kondition noch weniger bevorzugt. Als Gründe wurden mehrfach genannt, andere Museumsbesucher nicht stören zu wollen oder aufgrund der Umgebungslautstärke Schwierigkeiten mit der Spracherkennung oder Sprachausgabe zu befürchten.

Die hohe Präferenz für die textbasierte Kondition steht im Gegensatz zu der Erwartung, mit der Kombination von Spracheingabe und Textausgabe eine Alternative zu den symmetrischen Modalitätskombinationen gefunden zu haben. Als ausschlaggebender Faktor kann die hohe Benutzerkontrolle und Freiheit über die Textmodalität identifiziert werden. Zudem könnte die höhere Anzahl an Teilnehmenden mit Vorerfahrungen und häufigerer Nutzung chatbasierter Assistenzsysteme zu einer größeren Vertrautheit mit der textbasierten Interaktion geführt haben. Für die Konzeption zielgerichteter Dialogsysteme sollte daher eine textbasierte Interaktion als Grundlage bereitgestellt werden. Da ein hoher Task-Technology Fit benutzerabhängig ist [25] und Sprache eine effizientere Eingabemodalität darstellt, ist es sinnvoll, mehrere Interaktionsmodalitäten in Dialogsysteme zu integrieren.

3.6 Limitationen und Ausblick

Das gewählte Studiendesign ermöglichte einen sehr strengen Vergleich der Modalitäten und ihrer Kombinationsmöglichkeiten. Daraus ergeben sich einige Einschränkungen, die in zukünftigen Arbeiten berücksichtigt werden könnten.

Das in dieser Studie gewählte Frage-Antwort-Schema deckt nur den Bereich der zielgerichteten Dialogsysteme mit faktenbasierter Suche ab. Die Ergebnisse sind nicht auf die explorative Suche übertragbar, da in dieser Studie keine kontinuierliche Konversation zwischen Mensch und Maschine erzeugt wurde. Zukünftige Arbeiten könnten den Vergleich von Modalitätskombinationen auf die explorative Suche übertragen.

Die Analyse der Ergebnisse ergab zahlreiche Unterschiede, die jedoch nicht signifikant waren. Da nur jeweils drei kurze Fragen und Antworten in einer Kondition enthalten waren, ist es möglich, dass die Versuchspersonen zu wenig Eindrücke von der Modalitätskombination erhalten haben. Eine größere Stichprobe könnte ebenfalls helfen, diese Unterschiede zu verdeutlichen. Dazu wäre die Durchführung von Online-Studien denkbar, sofern die Funktionalität aller zu testenden Modalitäten sichergestellt werden kann. Auch wenn es schwierig wäre, gleiche Bedingungen zu gewährleisten, könnte die Anwendung so auf unterschiedlichen Benutzersystemen und unter realistischeren Rahmenbedingungen getestet werden. Die vorgestellten Szenarien bieten zusätzliche Anreize, die Studie als Feldexperiment mit angepassten Anwendungen zu wiederholen. Für das Szenario *Zuhause* bietet sich die Online-Studie an. Für das Szenario *Museum* könnte eine zugeschnittene Anwendung auf dem Smartphone oder ein virtueller Assistent auf einer Infostele im Museum getestet werden.

4 Zusammenfassung

Ziel dieser Arbeit war es, geeignete Ein- und Ausgabemodalitäten für ein zielgerichtetes Dialogsystem zur Informationsbeschaffung zu identifizieren. Dazu wurde eine Anwendung mit verschiedenen Ein- und Ausgabemodalitäten konzipiert und die daraus resultierenden Modalitätskombinationen mittels einer Benutzerstudie evaluiert. Ein Schwerpunkt lag dabei auf dem Vergleich asymmetrischer Kombinationen. Sowohl die symmetrische Spracheingabe mit Sprachausgabe als auch die asymmetrische Spracheingabe mit Textausgabe konnten in verschiedenen Kriterien signifikant bessere Ergebnisse erzielen. Die Spracheingabe führte zu einer signifikanten Verkürzung der Eingabezeit. Die Interaktion mit Sprache führte zu einer signifikant höheren wahrgenommenen hedonischen Qualität des Systems und zeigte eine höhere wahrgenommene Natürlichkeit. Dennoch wurde sowohl in der Laborbedingung als auch in allen vorgestellten Szenarien die Texteingabe mit Textausgabe als bevorzugte Modalitätskombination gewählt. Die rein textbasierte Kondition kann die höchsten Werte hinsichtlich Usability und pragmatischer Qualität erreichen, auch wenn sich diese nicht signifikant von den anderen Modalitätskombinationen unterscheiden. Eine bessere Erklärung liefern die Begründungen der Teilnehmenden, die das hohe Maß an Kontrolle über die textbasierte Interaktion hervorheben, was sich auch im geringsten Frustrationsniveau und Zeitdruckempfinden widerspiegelt. Die Benutzerkontrolle scheint somit einen stärkeren

Einfluss auf die Präferenz zu haben als die Effizienz oder die hedonische Qualität. Schließlich kann sich keine Modalitätskombination in allen Kriterien behaupten. Aufgrund der Präferenz für textbasierte Interaktionen sollte nach Möglichkeit sowohl eine textbasierte Eingabe als auch eine textbasierte Ausgabe zur Verfügung gestellt werden. Um einen optimalen Task-Technology Fit zu erreichen sollten zusätzliche Erweiterungen um eine sprachbasierte Eingabe, beispielsweise in Form eines Mikrofonsymbols neben dem Textfeld oder eines Aktivierungswortes, sowie einer sprachbasierten Ausgabe in Betracht gezogen werden. Diese Empfehlung richtet sich an die Gestaltung zielgerichteter Dialogsysteme. Im Hinblick auf künstliche Intelligenzen wird aber auch der nicht zielgerichtete Dialog weiter in den Fokus rücken. Zukünftige Arbeiten könnten das Studiendesign daher auf explorative Aufgabentypen übertragen oder in den genannten Szenarien validieren.

LITERATUR

[1] P. Y. Andrews. System Personality and Persuasion in Human-Computer Dialogue. *ACM Transactions on Interactive Intelligent Systems*, 2(2), Jun 2012. doi: 10.1145/2209310.2209315

[2] A. Atiyah, S. Jusoh, and F. Alghanim. Evaluation of the Naturalness of Chatbot Applications. In *IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pp. 359–365, 2019. doi: 10.1109/JEEIT.2019.8717455

[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, Jan 2009. doi: 10.1007/s12369-008-0001-3

[4] L. L. Bigot, J.-F. Rouet, and E. Jamet. Effects of Speech- and Text-Based Interaction Modes in Natural Language Human-Computer Dialogue. *Human Factors*, 49(6):1045–1053, 2007. doi: 10.1518/001872007X249901

[5] J. Brooke. SUS: A "Quick and Dirty" Usability Scale. *Usability Evaluation in Industry*, 189(3):189–194, 1996.

[6] B. R. Cowan et al. Understanding Speech and Language Interactions in HCI: The Importance of Theory-Based Human-Human Dialogue Research. In *Designing Speech and Language Interactions Workshop, ACM Conference on Human Factors in Computing Systems (CHI)*, 2014.

[7] F. L. I. Dutsinma, D. Pal, S. Funilkul, and J. H. Chan. A Systematic Review of Voice Assistant Usability: An ISO 9241–11 Approach. *SN Computer Science*, 3(4):267, May 2022. doi: 10.1007/s42979-022-01172-3

[8] U. Gnewuch, S. Morana, M. Adam, and A. Maedche. Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *European Conference on Information Systems (ECIS)*, 06 2018.

[9] D. L. Goodhue and R. L. Thompson. Task-Technology Fit and Individual Performance. *MIS Quarterly*, 19(2):213–236, 1995.

[10] B. Gordijn and H. t. Have. ChatGPT: Evolution or Revolution? *Medicine, Health Care and Philosophy*, 26(1):1–2, Mar 2023. doi: 10.1007/s11019-023-10136-0

[11] Y. Han, Y. Diao, Z. Yin, R. Jin, J. Kangwa, and O. J. Ebohon. Immersive Technology-Driven Investigations on Influence Factors of Cognitive Load Incurred in Construction Site Hazard Recognition, Analysis and Decision Making. *Advanced Engineering Informatics*, 48:101298, 2021. doi: 10.1016/j.aei.2021.101298

[12] S. G. Hart. NASA Task Load Index (TLX). 1986.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. doi: 10.1109/MSP.2012.2205597

[14] J. Hirschberg and C. D. Manning. Advances in Natural Language Processing. *Science*, 349(6245):261–266, 2015. doi: 10.1126/science.aaa8685

[15] W. J. King and J. Ohya. The Representation of Agents: Anthropomorphism, Agency, and Intelligence. In *ACM Conference Companion on Human Factors in Computing Systems (CHI)*, p. 289–290. New York, NY, USA, 1996. doi: 10.1145/257089.257326

[16] R. Kocielnik, D. Avrahami, J. Marlow, D. Lu, and G. Hsieh. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the ACM Designing Interactive Systems Conference (DIS)*, p. 881–894. New York, NY, USA, 2018. doi: 10.1145/3196709.3196784

[17] N. Kock. The Psychobiological Model: Towards a New Theory of Computer-Mediated Communication Based on Darwinian Evolution. *Organization Science*, 15(3):327–348, 2004. doi: 10.1287/orsc.1040.0071

[18] C. Krettek. ChatGPT. *Die Unfallchirurgie*, 126(3):252–254, Mar 2023. doi: 10.1007/s00113-023-01296-y

[19] C. Liebrecht, N. Kamoen, and C. Aerts. Voice Your Opinion! Young Voters' Usage and Perceptions of a Text-Based, Voice-Based and Text-Voice Combined Conversational Agent Voting Advice Application (CAVAA). In *Chatbot Research and Design*, pp. 34–49. Springer International Publishing, Cham, 2023. doi: 10.1007/978-3-031-25581-6_3

[20] M. Luria, G. Hoffman, and O. Zuckerman. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, p. 580–628. New York, NY, USA, 2017. doi: 10.1145/3025453.3025786

[21] X. Ma and A. Liu. Challenges in Supporting Exploratory Search through Voice Assistants. In *Proceedings of the ACM Conference on Conversational User Interfaces (CUI)*, pp. 1–3. New York, NY, USA, 2020. doi: 10.1145/3405755.3406152

[22] J. Nielsen. Ten Usability Heuristics. 2005.

[23] G. Norman. Likert Scales, Levels of Measurement and the "Laws" of Statistics. *Advances in Health Sciences Education*, 15(5):625–632, Dec 2010. doi: 10.1007/s10459-010-9222-y

[24] D. Pal, C. Arpnikanondt, S. Funilkul, and W. Chutimaskul. The Adoption Analysis of Voice-Based Smart IoT Products. *IEEE Internet of Things Journal*, 7(11):10852–10867, 2020. doi: 10.1109/JIOT.2020.2991791

[25] L. Riefle, A. Brand, J. Mietz, L. Rombach, C. Szekat, and C. Benz. What Fits Tim Might Not Fit Tom: Exploring the Impact of User Characteristics on Users' Experience with Conversational Interaction Modalities. *Wirtschaftsinformatik 2022 Proceedings*, 2022.

[26] S. Ruan, J. O. Wobbrock, K. Liou, A. Ng, and J. A. Landay. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4), Jan 2018. doi: 10.1145/3161187

[27] D. L. Rubin, T. Hafer, and K. Arata. Reading and Listening to Oral-Based Versus Literate-Based Discourse. *Communication Education*, 49(2):121–133, 2000. doi: 10.1080/03634520009379200

[28] C. Rzepka, B. Berger, and T. Hess. Voice Assistant vs. Chatbot – Examining the Fit Between Conversational Agents' Interaction Modalities and Information Search Tasks. *Information Systems Frontiers*, 24(3):839–856, Jun 2022. doi: 10.1007/s10796-021-10226-5

[29] N. Sawhney and C. Schmandt. Nomadic Radio: Speech and Audio Interaction for Contextual Messaging in Nomadic Environments. *ACM Transactions on Computer-Human Interaction*, 7(3):353–383, Sep 2000. doi: 10.1145/355324.355327

[30] A. Schmitt, N. Zierau, A. Janson, and J. M. Leimeister. Voice as a Contemporary Frontier of Interaction Design. In *European Conference on Information Systems (ECIS)*, 2021.

[31] M. Schrepp, A. Hinderks, and J. Thomaschewski. Konstruktion einer Kurzversion des User Experience Questionnaire. In *Mensch und Computer 2017 - Tagungsband*, pp. 355–360. Gesellschaft für Informatik e.V., Regensburg, 2017. doi: 10.18420/muc2017-mci-0006

[32] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *CoRR*, abs/1609.03499, 2016. doi: 10.48550/arXiv.1609.03499

[33] X. Wang and C. Yuan. Recent Advances on Human-Computer Dialogue. *CAAI Transactions on Intelligence Technology*, 1(4):303–312, 2016. doi: 10.1016/j.trit.2016.12.004

[34] N. V. Wünderlich and S. Paluch. A Nice and Friendly Chat with a Bot: User Perceptions of AI-Based Service Agents. In *International Conference on Interaction Sciences*, 2017.

A Verwendetes Goethe Aufgaben Material

Jahr der Ernennung Goethes zum Geheimen Rat: _____

Name Goethes erstgeborener Sohn: _____

Stadt, die Goethe in Italien besonders faszinierte: _____

Jahr der Veröffentlichung Goethes „Farbenlehre“: _____

Name eines berühmten Briefpartners Goethes: _____

Stadt, in der Goethe ein Ministeramt hatte: _____

Jahr Goethes erster Reise nach Karlsbad: _____

Name des Herzogs, der Goethe einen Adelstitel vergab: _____

Stadt, in der Goethe studierte: _____

Jahr der Veröffentlichung Goethes „Zauberlehrling“: _____

Name des ersten Goethe-Archiv Direktors: _____

Stadt, in der Goethe Napoleon traf: _____

Abbildung 4: Vollständige Frageblöcke zum Thema Goethe

Generiere mir einen Frageblock mit jeweils 3 kurzen Fragen zu Goethe und den dazugehörigen Antworten. Fragen und Antworten sollten alle ungefähr gleich lang sein. Der Inhalt soll weniger bekanntes Wissen über Goethe vermitteln. Die erste Frage soll nach einer Jahreszahl fragen. Die zweite Frage nach dem Namen einer Person in Goethes Leben. Die dritte Frage nach einem Ort in Goethes Leben. Die Antworten sollen in einem vollständigen Satz wiedergegeben werden.

Abbildung 5: ChatGPT Prompt zur Generierung der Frageblöcke

B Tabelle zur statistischen Auswertung

Tabelle 2: Mittelwerte und Standardabweichungen der erhobenen Daten für die jeweilige Kondition

	Text - Text	Text - Sprache	Sprache - Text	Sprache - Sprache
Anwendung (niedriger ist besser)				
Bearbeitungszeit (ms)	36614.54 (11294.08)	33087.18 (9062.02)	19257.83 (5834.02)	20307.73 (6587.17)
NASA-TLX (niedriger ist besser)				
Mental Demand	8.75 (15.79)	17.14 (21.36)	8.39 (10.81)	13.21 (13.35)
Physical Demand	8.75 (14.05)	10.36 (13.47)	5.18 (7.87)	3.21 (4.35)
Temporal Demand	9.11 (13.61)	15.71 (19.28)	10.54 (15.71)	11.79 (10.82)
Performance	7.32 (18.13)	7.14 (14.17)	8.04 (17.34)	7.86 (15.12)
Effort	8.04 (12.04)	13.04 (17.39)	8.39 (12.55)	8.39 (11.39)
Frustration	4.29 (6.63)	9.64 (16.66)	13.21 (18.01)	10.71 (12.23)
SUS (höher ist besser)				
Usability Score	89.29 (9.57)	82.41 (14.41)	87.05 (16.13)	87.41 (8.70)
UEQ-S (höher ist besser)				
Pragmatische Qualität	2.10 (0.85)	1.65 (1.27)	2.04 (1.04)	1.98 (0.85)
Hedonische Qualität	0.06 (1.25)	0.68 (1.21)	0.79 (1.21)	0.88 (1.06)
GQS (höher ist besser)				
Anthropomorphismus	2.65 (1.03)	2.70 (0.97)	2.71 (1.00)	3.10 (0.91)

C Weitere Bildschirmaufnahmen der mobilen Anwendung

Abbildung 6: Ankündigung der Eingabemodalität mit Countdown beim Start einer neuer Kondition

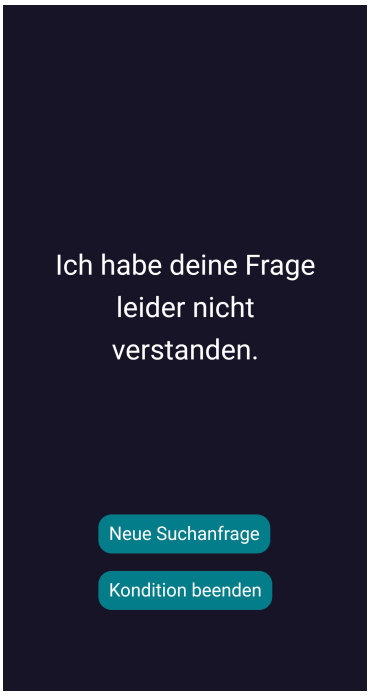


Abbildung 7: Darstellung des Fallback-Intents wenn das Intent Matching des Dialogflow Agenten fehlschlug

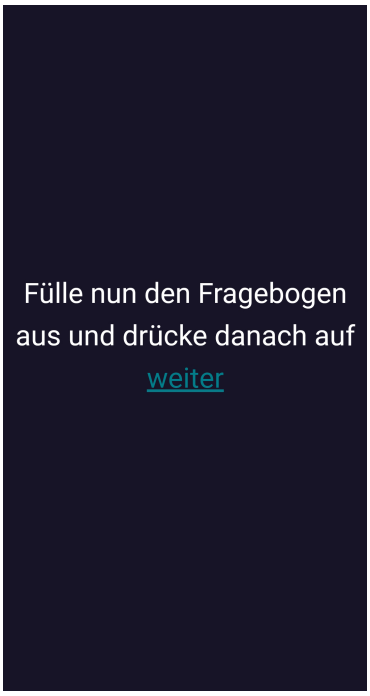


Abbildung 8: Hinweis zum Ausfüllen des Fragebogens nach dem Beenden einer Kondition

Ich bin damit einverstanden, dass meine Arbeit in den Bestand der Bibliothek eingestellt wird.

Ort, Datum	Unterschrift

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Studiengang Mensch-Computer-Interaktion selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel — insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen — benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum

Unterschrift