

MASTERARBEIT

Sprachinteraktion mit intelligenten virtuellen Agenten zur Fragebogenerhebung in Virtual Reality

vorgelegt am 4. November 2025
Mirco Hülsemann

Erstprüfer: Prof. Dr. Roland Greule
Zweitprüfer: Prof. Dr. Eike Langbehn

**HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG**
Fakultät Informatik und Digitale Gesellschaft
Finkenau 35
22081 Hamburg

Zusammenfassung

Die vorliegende Arbeit untersucht die sprachbasierte Fragebogenerhebung in Virtual Reality (VR) durch einen intelligenten virtuellen Agenten (IVA) als Alternative zu klassischen in-VR-Benutzeroberflächen (UI). Ziel war es, die technische Machbarkeit, die Datenäquivalenz der Antworten, das Nutzererlebnis und das Präsenzerleben beider Ansätze zu vergleichen. Dazu wurde eine experimentelle Within-Subjects-Studie mit 36 Teilnehmenden durchgeführt. Innerhalb von VR wurden auf einer Likert-Skala das Präsenzepfinden (IPQ) sowie drei Freitextfragen beantwortet – entweder per Sprachinteraktion mit einem LLM-gestützten IVA oder über eine klassische 2D-Benutzeroberfläche. Ergänzend wurden die wahrgenommene Arbeitsbelastung (NASA-TLX), die Gebrauchstauglichkeit (SUS), die Nutzererfahrung sowie technische Kennzahlen des Systems erhoben.

Die Ergebnisse zeigen, dass die IVA-Erhebung eine valide und praktikable Alternative zur klassischen UI-Befragung darstellt. Die quantitativen IPQ-Antworten und damit auch das subjektive Präsenzerleben blieben über beide Bedingungen hinweg äquivalent, während die sprachbasierten Freitextantworten signifikant länger und inhaltlich vielfältiger ausfielen. Die IVA-Interaktion wurde bei offenen Fragen als weniger belastend und gebrauchstauglicher erlebt, während bei Likert-Skalen-Fragen eine höhere Frustration auftrat, möglicherweise bedingt durch hohe Latenzen und repetitive Aussagen des Agenten.

Insgesamt zeigt die Studie, dass sprachgesteuerte IVAs eine verlässliche, nutzerfreundliche und datenäquivalente Methode zur Fragebogenerhebung in immersiven VR-Umgebungen bieten. Perspektivisch können lokal ausgeführte und weniger LLM-abhängige IVA-Systeme die Erhebung effizienter, datenschutzfreundlicher und barriereärmer gestalten und so eine bestehende methodische Lücke in der qualitativen Datenerfassung innerhalb von VR schließen.

Abstract

This thesis investigates speech-based questionnaire administration in Virtual Reality (VR) using an intelligent virtual agent (IVA) as an alternative to traditional in-VR user interfaces. The aim was to systematically compare the technical feasibility, data equivalence, user experience, and sense of presence of both approaches.

An experimental within-subjects study with 36 participants was conducted, in which, on a Likert scale, the sense of presence (IPQ) and three open-ended questions were answered within VR – either through spoken interaction with an LLM-powered IVA or via a conventional 2D user interface. In addition, workload (NASA-TLX), usability (SUS), user experience, and technical performance indicators were recorded.

The results indicate that the IVA-based survey represents a valid and practicable alternative to classical UI-based questionnaires. Quantitative IPQ responses and thus the reported sense of presence remained equivalent across both conditions, while spoken free-text answers were significantly longer and more diverse. Participants experienced the IVA condition as less demanding and more usable for open-ended questions, whereas higher frustration occurred with Likert-scale items, likely due to latency and repetitive agent utterances.

Overall, the study demonstrates that speech-driven IVAs provide a reliable, user-friendly, and data-equivalent method for questionnaire administration in immersive VR environments. In the future, locally executed and less LLM-dependent IVA systems could further improve efficiency, privacy, and accessibility, thereby addressing an existing methodological gap in qualitative data collection within VR.

Inhaltsverzeichnis

Abkürzungsverzeichnis	V
Abbildungsverzeichnis	VI
Tabellenverzeichnis	VIII
1 Einleitung	1
1.1 Zielsetzung und Forschungsfragen	2
1.2 Aufbau der Arbeit	3
2 Theoretischer Hintergrund	4
2.1 Virtual Reality, Immersion und Präsenz	4
2.1.1 Hardware und Interaktionsmethoden	5
2.1.2 Immersion und Präsenz	8
2.1.3 Messmethoden	10
2.2 Socially Interactive Agents	11
2.2.1 Darstellung und Verhalten	12
2.2.2 Gestik und Mimik	13
2.2.3 Soziale Präsenz und Rapport	14
2.2.4 Architekturen intelligenter Agenten	15
2.3 Neuronale Netze	17
2.3.1 Die Transformer-Architektur	18
2.3.2 Large Language Models	21
2.4 Sprachpipeline	23
2.4.1 Komponenten einer Sprachpipeline	24
2.4.2 Weitere Aspekte aus der Computerlinguistik	26
3 Aktueller Stand der Forschung	28
3.1 Fragebogenerhebungen in VR	28
3.2 Sprachinteraktion mit IVAs als Lösungsansatz	29
3.3 Herausforderungen LLM-basierter Agenten	31
3.4 Synthese und Positionierung der Forschung	33

4	Forschungsdesign und Methodik	35
4.1	Studiendesign	35
4.2	Versuchsablauf	36
4.3	Erhebungsinstrumente	37
4.4	Hypothesen und Operationalisierung	38
4.4.1	Hypothesen	39
4.4.2	Operationalisierung	40
4.5	Statistische Methoden	41
4.6	Pilotstudie	44
4.7	Versuchsaufbau	44
5	Umsetzung des Prototyps	46
5.1	Systemarchitektur	46
5.1.1	Zustandsmaschine	47
5.1.2	WebSocket-Kommunikation	49
5.2	VR-Szenario	50
5.3	IVA-Integration	52
5.3.1	Visuelles Design	52
5.3.2	Animationen	52
5.3.3	Aufbau der Befragung in VR	54
5.4	Sprachpipeline	55
5.4.1	Audioeingang	56
5.4.2	Voice Activity Detection	56
5.4.3	Automatic Speech Recognition	57
5.4.4	Natural Language Processing	58
5.4.5	Text-to-Speech	64
5.4.6	Audioausgabe	66
5.4.7	Fragebogen-Logik und Antwortverarbeitung	66
5.4.8	Turn-Taking	68
5.4.9	Modularität und Unabhängigkeit	69
5.4.10	Abgrenzung und nicht umgesetzte Funktionen	69
5.5	Fragebogen Benutzeroberfläche (2D-UI)	70
5.6	Protokollierung und Metrik-Framework	71
5.7	Entwicklungsumgebung und Hardware	72
6	Ergebnisse	74
6.1	Überblick und Stichprobe	74
6.2	Präsenzerleben (IPQ)	77
6.3	Freitextfragen zur VR-Erfahrung	80
6.3.1	Wortanzahl der Antworten	82

6.3.2	Anzahl kodierter Themen	83
6.3.3	Antworthäufigkeit und inhaltliche Qualität	84
6.4	Subjektive Arbeitsbelastung (NASA-TLX)	85
6.5	System Usability (SUS)	87
6.6	Reliabilität der Erhebungsinstrumente	88
6.7	Ergebnisse des Abschlussfragebogens	89
6.7.1	Gesamtgefühl	90
6.7.2	Vertraulichkeit der Antworten	91
6.7.3	Vertrauenswürdigkeit des IVA	92
6.7.4	Verständlichkeit der Fragen	93
6.7.5	Beantworten von Likert-Fragen	93
6.7.6	Beantworten von Freitextfragen	94
6.7.7	Auftreten von Schwierigkeiten	94
6.7.8	Präferenzen zu Erscheinungsbild und Aussehen	96
6.7.9	Präferenzen für Antwortformate	98
6.7.10	Verbesserungsvorschläge	101
6.8	Technische Kennzahlen und Leistungsanalyse	103
6.8.1	Latenzen der Sprachpipeline	103
6.8.2	Nutzerverhalten und Interaktionsdynamik	105
6.8.3	Fehlerrate der Spracherkennung (ASR)	106
6.8.4	Erkennung von Sprachaktivität (VAD)	107
6.8.5	Fehleranalyse	107
6.9	Zusammenhänge und Einflussfaktoren	109
7	Diskussion	112
7.1	Zusammenfassung der Ergebnisse	112
7.1.1	Zuverlässigkeit	112
7.1.2	Datenäquivalenz und Antwortqualität	113
7.1.3	Nutzererlebnis	113
7.1.4	Präsenzerleben	114
7.2	Interpretation der Ergebnisse	115
7.2.1	H1 – Technische und subjektive Zuverlässigkeit	115
7.2.2	H2 – Datenäquivalenz und Antwortqualität	117
7.2.3	H3 – Nutzererfahrung	118
7.2.4	H4 – Präsenzerleben	120
7.3	Limitationen und Ausblick	121
7.4	Ausblick	122
8	Fazit	124

Literatur	127
Benutzte Hilfsmittel	138
Anhang	139
A LLM Prompts	139
A.1 Systeminstruktionen	139
A.2 Ausgabeformat	143
B Erhebungsinstrumente (Fragebögen)	144
B.1 In-VR-Fragebögen	144
B.1.1 Igroup Presence Questionnaire (IPQ)	144
B.1.2 Offene Fragen	145
B.2 Außerhalb-von-VR-Fragebögen	145
B.2.1 NASA Task Load Index (Raw NASA-TLX)	145
B.2.2 System Usability Scale (SUS)	146
B.2.3 Post-Survey (Abschlussfragebogen)	146
B.2.4 Demografischer Fragebogen	148
C Tabellen	149
C.1 Interne Konsistenz	149
C.2 Korrelationen	150

Abkürzungsverzeichnis

API	Application Programming Interface
ASR	Automatic Speech Recognition
BIP	Break in Presence
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
ECA	Embodied Conversational Agents
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HMD	Head-Mounted Display
inVRQ	in-VR-Questionnaire
ID	Identifikationsnummer
IPQ	Igroup Presence Questionnaire
IVA	Intelligent Virtual Agent
JSON	JavaScript Object Notation
KI	Künstliche Intelligenz
LLM	Large Language Model
LSTM	Long Short-Term Memory
LM	Language Model
NASA-TLX	NASA Task Load Index
NLP	Natural Language Processing
outVRQ	out-of-VR-Questionnaire
RNN	Recurrent Neural Network
SUS	System Usability Scale
TTS	Text-to-Speech
UI	User Interface
UX	User Experience
VAD	Voice Activity Detection
VR	Virtual Reality
WER	Word Error Rate
XR	Extended Reality

Abbildungsverzeichnis

2.1	Das Reality-Virtuality-Continuum	5
2.2	Meta Quest 3: ein aktuelles Standalone-VR-Headset mit Inside-Out-Tracking und Hand-Tracking	6
2.3	Schematische Darstellung typischer VR-Eingabemethoden	6
2.4	Schematische Darstellung dreier Interaktionsformen in VR anhand von Farbauswahl als Beispiel	7
2.5	Verschiedene Darstellungsweisen von Agenten	12
2.6	Darstellung des „Uncanny Valley“-Konzepts	13
2.7	Verarbeitungspfade multimodalen Verhaltens in IVA-Systemen	15
2.8	Schematische Darstellung eines Recurrent Neural Network (RNN)	18
2.9	Schematische Darstellung einer vollständigen Transformer-Architektur	19
2.10	Schematische Darstellung der konventionellen Sprachpipeline	23
3.1	Beispiele für verschiedene Darstellungsweisen von inVRQs	29
3.2	Visuelle Gestaltung von IVAs in der Literatur	31
3.3	Gegenüberstellung von vier Strategien zum Umgang mit Systemlatenz bei virtuellen Agenten	33
4.1	Darstellung des Versuchsaufbaus	45
5.1	Übersicht der Systemarchitektur des Prototyps	47
5.2	Zustandsdiagramm der Zustandsmaschine	48
5.3	Screenshots des VR-Szenarios in Unity	50
5.4	Interaktions- und Navigationsmechanismen im VR-Szenario	51
5.5	3D-Modell Kevin von Reallusion	52
5.6	Screenshots aus Animationen zu den jeweiligen Systemzuständen	53
5.7	Visualisierung der über Blendshapes erzeugten Mundbilder zu den Lauten (Phonemen) ah, eh und uh (woo)	53
5.8	Screenshots der Befragungsszene vor dem Laborzelt in VR	54
5.9	Module der der Sprachpipeline mit markiertem Audioeingang	56
5.10	Module der Sprachpipeline mit markierter Voice Activity Detection	56
5.11	Module der Sprachpipeline mit markierter Automatic Speech Recognition	57
5.12	Module der Sprachpipeline mit markiertem Natural Language Processing	58

5.13	Struktur des LLM-Requests	59
5.14	Module der Sprachpipeline mit markiertem Text-to-Speech-Modul	64
5.15	Module der Sprachpipeline mit markiertem Audioausgang	66
5.16	Screenshots des Virtual Reality Questionnaire Toolkits	71
6.1	Boxplots und Einzelwerte der Durchführungszeiten für VR-Erfahrung und Fragebogenerhebung	76
6.2	Violinplots der Verteilungen der IPQ-Subskalen- und Gesamtscores	77
6.3	Boxplots der IPQ-Subskalen- und Gesamtscores (14 Items)	79
6.4	Wordclouds der kodierten Themen der Freitextantworten, zusammengeführt für alle Fragen	81
6.5	Wortanzahl der Freitextantworten in den offenen Fragen des inVRQ je Bedingung	82
6.6	Anzahl kodierter Themen in den offenen Fragen	83
6.7	Verteilung der NASA-TLX-Werte für Likert-Skalen und Freitextfragen in beiden Bedingungen	85
6.8	System Usability Scale (SUS)-Scores für IVA und UI	87
6.9	Antwortverteilungen im Abschlussfragebogen	89
6.10	Präferenzen der Teilnehmenden für IVA vs. UI in Abhängigkeit vom Fragetyp	99
6.11	Vergleich der Latenzen über die Verarbeitungsstufen hinweg (log-Skala)	103
6.12	Histogramm der Gesamt-Pipeline-Latenzen	104
6.13	Verteilung der User-Latenzen (Zeit bis zur nächsten Nutzereingabe)	105
6.14	Korrelationsmatrix zwischen den zentralen Skalen (Spearman- ρ)	110

Tabellenverzeichnis

5.1	Ausgabeschema der Modellantworten im Structured Output Format	60
5.2	Überblick über Entwicklungsumgebung und Hardware	73
6.1	Erfahrung der Teilnehmenden mit Virtual Reality	74
6.2	Erfahrung der Teilnehmenden mit Sprachassistenten	75
6.3	Dauer der VR-Erfahrungen und Fragebogenerhebungen	76
6.4	Dispersion und gepaarter Varianzvergleich (IPQ)	78
6.5	Schiefe γ und Schiefe-Differenzen der IPQ-Scores zwischen Erhebungsformen	78
6.6	Mittelwerte und Standardabweichungen für das subjektive Präsenzerleben (IPQ)	80
6.7	Häufigste kodierte Themen (Gesamt und pro Bedingung)	81
6.8	Wortanzahl in den Freitextantworten pro Antwort	82
6.9	Anzahl kodierter Themen in den Freitextantworten pro Antwort	84
6.10	Antworthäufigkeiten und Anteil inhaltlicher Antworten pro Bedingung und Frage	84
6.11	Ergebnisse des Raw NASA-TLX pro Bedingung für Likert-Skalen und Frei- textfragen	86
6.12	System Usability Scale (SUS) in beiden Bedingungen	87
6.13	Interne Konsistenz (Cronbach's α) der eingesetzten Fragebögen in beiden Bedingungen (Gesamtwerte)	88
6.14	Mittelwerte der Antworten im Abschlussfragebogen	90
6.15	Auftreten von Schwierigkeiten während der Befragung	95
6.16	Präferenzen für das Erscheinungsbild des KI-Agenten	96
6.17	Präferenzen für IVA vs. UI in Abhängigkeit vom Fragetyp	98
6.18	Latenzen der Sprachpipeline	104
6.19	Auftreten von Fehlermeldungen in den Gesprächsrunden	107
C.1	Interne Konsistenz (Cronbach's α) der eingesetzten Skalen und Subskalen . .	149
C.2	Korrelationen (Spearman's ρ) zwischen demografischen Variablen, psycho- metrischen Skalen und Präferenzmaßen	150

1 Einleitung

Virtual Reality (VR) hat sich von einer Nischentechnologie der Unterhaltungsbranche zu einem etablierten Werkzeug in Forschung, Therapie und Bildung entwickelt. Die Fähigkeit von VR, Nutzer in computergenerierte Welten eintauchen zu lassen und ein Gefühl der tatsächlichen Anwesenheit – der sogenannten Präsenz (Sherman & Craig, 2002) – zu erzeugen, eröffnet einzigartige Möglichkeiten für realitätsnahe Simulationen und Studien (Tepe et al., 2024). Doch gerade diese immersive Stärke stellt die empirische Forschung vor eine methodische Herausforderung: Wie können Daten, insbesondere subjektives Feedback, erfasst werden, ohne das Gefühl der Präsenz zu brechen und die Qualität der Ergebnisse zu beeinträchtigen?

Traditionell erfolgt die Beantwortung von Fragebögen nach einer VR-Erfahrung außerhalb der virtuellen Realität auf Papier oder digital. Dieser Medienwechsel zwingt die Teilnehmenden, das Head-Mounted Display (HMD) abzulegen, was zu einem abrupten Bruch der Präsenzerfahrung führt, einem sogenannten „Break in Presence“ (BIP) (Slater & Steed, 2000). Solche Unterbrechungen können nicht nur zu Desorientierung und Verunsicherung führen (Knibbe et al., 2018), sondern auch die Ergebnisse subjektiver Selbstberichte beeinträchtigen (Putze et al., 2020; Schwind et al., 2019).

Um dieses Problem zu umgehen, werden Fragebögen vermehrt direkt in die VR-Umgebung integriert (inVRQ). Studien zeigen, dass dieser Ansatz die Konsistenz der Antworten erhöht, ohne die Mittelwerte im Vergleich zu traditionellen Methoden zu verändern (Schwind et al., 2019). Zudem werden In-VR-Befragungen von Nutzenden oft als ansprechender und natürlicher Teil des VR-Erlebnisses wahrgenommen (Alexandrovsky et al., 2020; Safikhani et al., 2021).

Trotz dieser Vorteile bleibt die Eingabe über virtuelle Benutzeroberflächen eine Herausforderung. Insbesondere die Erfassung von offenen Freitextantworten stellt eine Hürde dar. Das Tippen auf virtuellen Tastaturen ist im Vergleich zu physischen Pendanten deutlich langsamer und kognitiv belastender (Alexandrovsky et al., 2021; Knierim et al., 2018). Da offene Fragen für qualitatives Feedback jedoch unverzichtbar sind (Alexandrovsky et al., 2021), besteht hier ein Forschungsanliegen.

An dieser Stelle setzt die vorliegende Arbeit an und untersucht einen neuartigen Ansatz, der die Nachteile klassischer Benutzeroberflächen in VR umgeht: die sprachbasierte Befragung

durch einen Intelligent Virtual Agent (IVA) direkt in VR. Moderne Sprachmodelle (LLMs) ermöglichen heute natürliche Dialoge, die weit über einfache Sprachbefehle hinausgehen (Wang et al., 2025). Während Sprachinteraktion mit IVAs in anderen Anwendungskontexten bereits untersucht wurde, fehlt bislang eine systematische Untersuchung der Fragebogenerhebung durch sprachbasierte virtuelle Agenten in VR. Die vorliegende Arbeit adressiert damit eine bislang offene methodische Lücke. Die Befragung durch einen virtuellen Charakter könnte die Interaktion außerdem insgesamt natürlicher und motivierender gestalten (Soni et al., 2022) und eröffnet zugleich neue Möglichkeiten der Barrierefreiheit, indem auch Personen mit motorischen Einschränkungen einbezogen werden können (Mott et al., 2020).

1.1 Zielsetzung und Forschungsfragen

Das Ziel dieser Arbeit ist die Konzeption, Implementierung und Evaluation eines Systems zur automatisierten sprachbasierten Fragebogenerhebung durch einen KI-gesteuerten Agenten in Virtual Reality. Verglichen wird dieser Ansatz mit der etablierten Methode der In-VR-Fragebögen über eine klassische Benutzeroberfläche (UI). Im Mittelpunkt steht die Frage, ob ein sprachgesteuerter IVA eine zuverlässige, valide und nutzerfreundliche Alternative zur klassischen Eingabeform darstellt, sowohl für die Erfassung von quantitativen als auch von qualitativen Daten. Ebenso werden Präferenzen und subjektive Eindrücke der Teilnehmenden erhoben, um ein umfassendes Bild der Nutzererfahrung zu gewinnen.

Aus dieser Zielsetzung ergeben sich vier wesentliche Forschungsfragen:

1. **Zuverlässigkeit:** Wie zuverlässig funktioniert die Beantwortung von Fragebögen durch einen IVA in VR aus technischer und subjektiver Perspektive?
2. **Datenäquivalenz und Antwortqualität:** Unterscheiden sich die gegebenen Antworten auf Fragebögen, wenn diese in VR durch einen IVA oder über eine klassische 2D-UI ausgefüllt werden?
3. **Nutzererlebnis:** Wie wird die Nutzererfahrung (UX) bei der Befragung durch einen IVA im Vergleich zur Interaktion über eine klassischen 2D-UI wahrgenommen?
4. **Präsenzerleben:** Wie wirkt sich die Befragung durch einen IVA im Vergleich zur Interaktion über eine 2D-UI auf das subjektive Präsenzerleben in der VR-Umgebung aus?

Zur Beantwortung dieser Fragen wird eine experimentelle Studie durchgeführt, bei der alle Teilnehmenden beide Erhebungsformen durchlaufen: die Befragung durch einen sprachbasierten IVA sowie die Beantwortung über eine klassische 2D-UI in VR.

1.2 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich wie folgt: In **Kapitel 1** werden die Zielsetzung der Arbeit, die zugrunde liegenden Forschungsfragen sowie die formulierten Hypothesen vorgestellt. **Kapitel 2** führt in die theoretischen Grundlagen ein und behandelt Konzepte zu Virtual Reality, Fragebogenerhebung in VR, intelligenten virtuellen Agenten sowie den technischen Komponenten einer Sprachpipeline. Der aktuelle Forschungsstand wird in **Kapitel 3** dargestellt, indem es relevante Studien zusammenfasst und daraus die zugrunde liegende Forschungslücke ableitet. Das Forschungsdesign und die Methodik der Nutzerstudie, einschließlich Studiendesign, Ablaufplan, VR-Szenario, eingesetzter Technologien und Messinstrumente werden in **Kapitel 4** beschrieben. In **Kapitel 5** wird die Umsetzung des entwickelten Prototyps mit seinen Systemkomponenten dokumentiert. Die Ergebnisse der Studie werden in **Kapitel 6** präsentiert und in **Kapitel 7** diskutiert. **Kapitel 8** beinhaltet das Fazit, die Zusammenfassung der wichtigsten Erkenntnisse, die Benennung der Limitationen und schließt mit einem Ausblick auf zukünftige Forschung.

2 Theoretischer Hintergrund

Das folgende Kapitel legt die theoretischen und technologischen Grundlagen für das Verständnis der Arbeit. Zunächst wird der Kontext der virtuellen Realität eingeführt. Zentrale Begriffe wie Immersion und Präsenz, die für das Erleben virtueller Umgebungen entscheidend sind, werden erläutert. Zudem wird ein Überblick über die notwendige Hardware, wie Head-Mounted Displays, sowie gängige Interaktionsmethoden, einschließlich Controller- und Sprachsteuerung, gegeben.

Anschließend wird der Begriff des intelligenten virtuellen Agenten definiert und von verwandten Konzepten wie Avataren und Embodied Conversational Agents abgegrenzt. Dabei werden auch verschiedene Darstellungs- und Verhaltensweisen sowie technische Architekturen von IVA-Systemen beschrieben.

Es folgt eine Einführung in die Grundlagen von neuronalen Netzen, der sogenannten Künstlichen Intelligenz (KI). Danach werden Transformermodelle als moderne Architektur für die Verarbeitung sequenzieller Daten und darauf aufbauend Large Language Models (LLMs) vorgestellt.

Abschließend wird die technische Basis der Sprachinteraktion in Form einer typischen Sprachpipeline erläutert. Diese umfasst die drei Kernkomponenten der automatischen Spracherkennung (ASR), der Verarbeitung natürlicher Sprache (NLP) und der Text-zu-Sprache-Synthese (TTS).

2.1 Virtual Reality, Immersion und Präsenz

Virtual Reality (VR) bezeichnet eine computergenerierte, dreidimensionale Simulationsumgebung, in der Benutzer:innen realitätsnahe Erfahrungen machen und mit virtuellen Objekten interagieren können. Nach Tepe et al. (2024) umfasst VR interaktive und nicht-interaktive Simulationsanwendungen, die mithilfe tragbarer Geräte und bildgenerierender Technologien ein Gefühl der Anwesenheit in einer virtuellen Welt erzeugen. Um das Gehirn davon zu überzeugen, dass die künstliche Umgebung real ist, passt sich die Simulation den Bewegungen der Benutzer:innen an, sodass der Eindruck entsteht, tatsächlich in die virtuelle Welt einzutauchen („immersed“) oder sich darin anwesend („present“) zu fühlen (Craig et al., 2009).

Ein hilfreiches konzeptuelles Modell zur Einordnung virtueller Erfahrungen ist das Reality-Virtuality-Continuum von Milgram et al. (1994). Es beschreibt ein Spektrum zwischen vollständig realen und vollständig virtuellen Umgebungen, in dem auch Zwischenformen wie (AR) oder (AV) verortet werden (siehe Abb. 2.1). Der heute gebräuchliche Oberbegriff Extended Reality (XR) umfasst dieses gesamte Kontinuum. Die in dieser Arbeit untersuchte Virtual-Reality-Anwendung befindet sich am Ende des Spektrums in der vollständig virtuellen Realität.

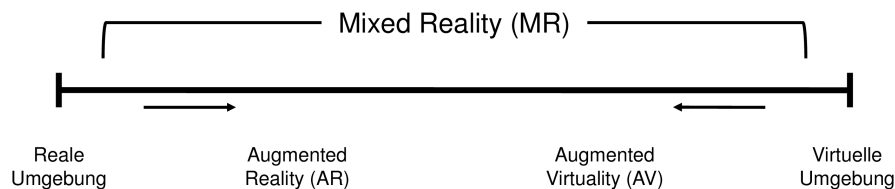


Abbildung 2.1: Reality-Virtuality-Continuum mit Spektrum zwischen realer und virtueller Umgebung (adaptiert aus Milgram et al., 1994)

2.1.1 Hardware und Interaktionsmethoden

Um immersive Erfahrungen in Virtual Reality zu ermöglichen, ist spezifische Hardware erforderlich. Das sogenannte HMD (Head-Mounted Display), ein am Kopf getragenes Gerät, erzeugt stereoskopische Bilder direkt vor den Augen der Nutzenden. Mithilfe von internen (sogenanntes Inside-Out-Tracking) oder externen (Outside-In-Tracking) Bewegungssensoren werden Kopf-, Hand- und teilweise auch andere Körperbewegungen erfasst und in Echtzeit in die virtuelle Welt übertragen.

Moderne VR-Systeme sind mit leistungsstarker Grafik- und Rechenleistung, Stereodisplays mit großem Sichtfeld und sechs Freiheitsgraden für die Kopfverfolgung sowie teilweise auch für die Verfolgung von Händen und Füßen ausgestattet (Slater, 2018). Einige aktuelle HMDs bieten zudem Passthrough, also die Anzeige der realen Umgebung über Kameras im Headset, wodurch unter anderem Mixed-Reality-Anwendungen möglich werden. Inzwischen sind am Markt auch sogenannte Standalone-VR-Headsets verfügbar, die keine Verbindung zu einem PC erfordern. Der Preis moderner HMDs liegt zwischen wenigen hundert Euro und einem niedrigen vierstelligen Betrag für Systeme im oberen Leistungs- und Preissegment, die sich durch Eigenschaften wie höhere Displayauflösung, gesteigerte Bildwiederholrate, größeres Sichtfeld sowie präzisere Trackingfunktionen auszeichnen (Schuir et al., 2022).



Abbildung 2.2: Meta Quest 3 (Meta Platforms, Inc., 2023): ein aktuelles Standalone-VR-Headset mit Inside-Out-Tracking und Hand-Tracking. (Eigene Darstellung)

Die Meta Quest 3 (siehe Abb. 2.2) ist ein aktuelles Standalone-VR-Headset, mit Inside-Out-Tracking, kamerabasiertem Hand-Tracking sowie Passthrough. Die Produktreihe von Meta wird häufig als technologisch ausgereift und zugleich preislich attraktiv beschrieben, steht jedoch aufgrund datenschutzbezogener Aspekte auch in der Kritik (Schuir et al., 2022).

Für die Interaktion in VR kommen verschiedene Eingabemethoden zum Einsatz. Am weitesten verbreitet sind VR-Controller, die Bewegungen im Raum erfassen und Tasten- oder Gesteneingaben ermöglichen. Alternativ oder ergänzend kommt zunehmend kamerabasiertes Hand-Tracking zum Einsatz, bei dem die Bewegungen und Gesten der Hände ohne physische Geräte erfasst werden. Weitere Formen der Eingabe sind die Sprachsteuerung, bei der Nutzer:innen über gesprochene Befehle mit der virtuellen Umgebung interagieren können, und Blicksteuerung (Gaze), bei der die Blickrichtung der Nutzer:innen erfasst wird, um Interaktionen auszulösen. Abbildung 2.3 zeigt eine Übersicht der genannten Eingabemethoden.

Diese Eingabemethoden bieten ein breites Spektrum an Interaktionsmöglichkeiten, unterscheiden sich jedoch in Bezug auf Präzision, Komfort und Einsatzszenarien. Ihre Eignung hängt stark vom Nutzungskontext ab und davon, ob sie einzeln oder kombiniert eingesetzt werden (Wang et al., 2025).



Abbildung 2.3: Schematische Darstellung typischer Eingabemethoden in VR (v.l.n.r.): VR-Controller, kamerabasiertes Hand-Tracking, Sprachsteuerung und Blicksteuerung (Illustration generiert von ChatGPT)

Controller und handbasiertes Tracking ermöglichen eine präzise und zuverlässige Steuerung in XR-Umgebungen, etwa bei der Selektion oder Manipulation von Objekten. Eine weitverbreitete Technik für die Interaktion mittels VR-Controllern ist das sogenannte Raycasting. Dabei wird ein Strahl – vergleichbar mit einem Laserpointer – vom Controller in die virtuelle Umgebung projiziert. Trifft dieser Strahl auf ein interaktives Objekt, kann mit diesem durch eine Eingabe am Controller interagiert werden. Diese Methode ermöglicht eine präzise Selektion von Elementen auch über größere Entfernungen (Pietroszek, 2018). Allerdings erfordert die Nutzung von Controllern gezielte Bewegungen über einen längeren Zeitraum, was zu einer Ermüdung der Arme führen kann.

Sprachsteuerung bietet eine freihändige und körperlich weniger beanspruchende Alternative. Besonders bei Aufgaben, die keine exakten räumlichen Eingaben erfordern, kann Sprache eine intuitive Form der Interaktion darstellen. Gleichzeitig ist sie anfällig für Erkennungsfehler und Verzögerungen, insbesondere in geräuschvollen Umgebungen, und kann durch die Notwendigkeit, bestimmte Befehle zu erinnern, kognitiv fordernd sein. Auch soziale Hemmnisse, etwa beim Sprechen in öffentlichen Räumen, stellen eine Herausforderung dar (Wang et al., 2025).

Eine vergleichende Studie von Hepperle et al. (2019) zeigt, dass Spracheingabe am leichtesten zu erlernen und besonders effizient für die Texteingabe ist, jedoch mit einer höheren Fehleranfälligkeit einhergeht. Während 2D-Interfaces die größte Präzision bieten, werden 3D-Interaktionen als immersiver wahrgenommen, können aber auch schwerer zu erlernen sein. Laut den Autoren ist die Eignung einer Methode sowohl stark aufgabenspezifisch als auch von der Vorerfahrung und Voreingenommenheit der Nutzer abhängig.

Abbildung 2.4 veranschaulicht diese drei grundlegenden Interaktionsformen exemplarisch anhand einer Farbauswahlaufgabe in Virtual Reality.



Abbildung 2.4: Schematische Darstellung dreier Interaktionsformen in VR anhand von Farbauswahl als Beispiel (Hepperle et al., 2019). 2D-UI über Menüs (links), räumliche 3D-Interaktion mit virtuellen Objekten (Mitte) und sprachbasierte Interaktion (rechts) (Illustration generiert von ChatGPT, eigene Darstellung)

Fortschritte bei Large Language Models (LLMs) haben die reine Sprachsteuerung erheblich weiterentwickelt. Wie Wang et al., 2025 beschreiben, waren frühere Systeme oft auf feste Schlüsselwörter beschränkt, wohingegen moderne Ansätze offene, natürliche Dialoge ermöglichen. Dabei wird Sprache zunehmend mit anderen Modalitäten kombiniert, etwa mit Blick- oder Gestensteuerung. So kann Sprache die semantische Bedeutung einer Anfrage liefern (z. B. „Was ist das?“), während die Blickrichtung oder eine Zeigegeste das gemeinte Objekt eindeutig identifiziert (Wang et al., 2025). Diese multimodale Verarbeitung macht Sprachinteraktion zu einer flexiblen und menschenähnlichen Schnittstelle für XR-Anwendungen.

2.1.2 Immersion und Präsenz

In dem Zusammenhang von VR werden häufig die Begriffe *Immersion* und *Presence* verwendet (z. B. Sherman und Craig, 2002; Slater, 2018; Tepe et al., 2024). Beide Begriffe sind jedoch nicht eindeutig und werden in der Literatur häufig unterschiedlich definiert oder sogar gleichgesetzt (Berkman & Akan, 2024). Dennoch dienen sie häufig als Messgrößen für die Erforschung von VR (Barešytė et al., 2024).

Immersion ist eng mit der technischen Ausstattung und Gestaltung eines VR-Systems verbunden. Slater und Wilbur (1997) beschreiben sie als die Fähigkeit eines Systems, eine lebendige, umschließende und glaubhafte Illusion für die menschlichen Sinne zu erzeugen. Andere Ansätze differenzieren zwischen sensorischer Immersion (z. B. audiovisuelle Reize, Steuerung, Interaktivität) und narrativer Immersion (inhaltliche und emotionale Ebene) (Botella et al., 2005; Ermi & Mäyrä, 2005). Witmer und Singer (1998) verstehen den Begriff hingegen als einen psychologischen Zustand der vollständigen Einbindung in eine Umgebung. Aufbauend darauf definieren Berkman und Akan (2024) Immersion als den objektiven Grad an sensorischer Genauigkeit, den ein VR-System bietet.

Presence (*Präsenz*) bezeichnet das psychologische Phänomen des Empfindens, sich tatsächlich in der virtuellen Umgebung zu befinden. Dieses Gefühl wird auch als „sense of being there“ beschrieben (Witmer & Singer, 1998) und entsteht meist unbewusst aus automatischen Reaktionen des sensorischen und motorischen Systems auf virtuelle Reize, auch wenn den Nutzenden bewusst ist, dass es sich um eine Simulation handelt (Slater, 2018). Presence ist nicht ausschließlich an VR gebunden, tritt dort jedoch besonders stark auf. Sie wird nicht nur von technischen Eigenschaften beeinflusst, sondern auch von individuellen Faktoren wie Persönlichkeitsmerkmalen, Immersionsneigung oder Aufmerksamkeit (Berkman & Akan, 2024).

Zusammenhang von Immersion und Presence

Obwohl Immersion und Presence eng verknüpft sind, beschreiben sie unterschiedliche Konzepte. Immersion verweist auf die objektiven Eigenschaften eines Systems, während Presence die subjektive Reaktion der Nutzenden darauf darstellt (Berkman & Akan, 2024; Slater, 2003). Ein hoch immersives System kann die Entstehung von Presence stark begünstigen, garantiert sie jedoch nicht zwangsläufig. Vielmehr entsteht Presence aus dem Zusammenspiel technischer Immersion und der individuellen psychologischen Reaktion der Nutzenden (Diemer et al., 2015; Slater, 2003). Presence ist somit ein eigenständiges psychologisches Phänomen, das aus der Immersion hervorgeht.

Embodiment, Soziale Präsenz und Rapport

Embodiment beschreibt allgemein das Erleben, sich mit einem virtuellen Körper zu identifizieren. Es entsteht durch die Wahrnehmung, dass der eigene Avatar den physischen Körper repräsentiert und dessen Bewegungen oder Handlungen in der virtuellen Umgebung widerspiegelt (Fribourg et al., 2020; Rzeszewski et al., 2024).

Social Presence beschreibt hingegen das Gefühl, die Anwesenheit anderer Akteure in einer virtuellen Umgebung wahrzunehmen und mit ihnen in Beziehung zu stehen. Moderne Ansätze differenzieren Social Presence in Dimensionen wie Co-Präsenz (das Gefühl, die Anwesenheit und Aufmerksamkeit des Gegenübers zu spüren), affektives Verstehen (das Erkennen und Verstehen der Emotionen des anderen) und Verhaltensinterdependenz (die gegenseitige Beeinflussung der Handlungen) (Harms & Biocca, 2004). Gerade in VR wird Social Presence durch Avatare verstärkt, da sie verkörperte Interaktion und soziale Nähe vermitteln können (Rzeszewski et al., 2024).

Eng damit verbunden ist der Begriff des Rapport, der die Qualität der sozialen Beziehung zwischen Interaktionspartnern beschreibt. Rapport umfasst Faktoren wie Vertrauen, Sympathie, Aufmerksamkeit und die Wahrnehmung eines harmonischen Gesprächsflusses. In der Forschung zu virtuellen Agenten wird er als entscheidender Faktor angesehen, um eine glaubwürdige, angenehme und kooperative Interaktion zu ermöglichen (Gratch & Lucas, 2021). Während Social Presence primär die Wahrnehmung der Anwesenheit eines Gegenübers beschreibt, geht Rapport darüber hinaus und erfasst die Qualität und Tiefe der sozialen Bindung, die während der Interaktion entsteht. Empirische Studien zeigen zudem, dass auch virtuelle Avatare und Agenten Rapport aufbauen können: Durch gezielte verbale und nonverbale Verhaltensweisen schaffen sie Vertrauen und erhöhen die Bereitschaft zur Selbstoffenbarung, insbesondere dann, wenn Teilnehmende das Gefühl haben, weniger stark bewertet zu werden (Lucas et al., 2014).

2.1.3 Messmethoden

Zur Evaluation von VR-Erfahrungen werden sowohl explizite als auch implizite Verfahren eingesetzt. Explizite Methoden wie Fragebögen und Interviews setzen eine aktive Beteiligung der Teilnehmenden voraus, während bei impliziten Verfahren (wie etwa der Blickbewegungsanalyse, der Herzratenvariabilität oder dem EEG) unbewusste Reaktionen erfasst werden. In der VR-Forschung werden Fragebögen besonders häufig verwendet, da sie sich leicht einsetzen lassen, thematisch vielseitig sind und die Auswertung subjektiver Erfahrungen ermöglichen (Bareišytė et al., 2024).

In der Literatur findet sich eine Vielzahl unterschiedlicher Fragebögen. Die am häufigsten vertretenen Schwerpunkte sind Präsenz, User Experience (UX), mentale Arbeitsbelastung (*mental workload*) oder VR-Krankheit (*motion sickness*). Präsenz wird dabei häufig in die Kategorien räumliche (*spatial-*), soziale (*social-*) und Selbstpräsenz (*self-presence*) unterteilt (Bareišytė et al., 2024; Safikhani et al., 2024). Derzeit mangelt es an allgemein anerkannten Standards zur Auswahl geeigneter Fragebögen. Bareišytė et al. (2024) liefern jedoch erstmals eine systematische Empfehlung für unterschiedliche Messziele in VR-Kontexten, differenziert nach kompakten (<15 Items) und ausführlicheren Instrumenten. Beispielsweise empfehlen sie den IPQ (Schubert et al., 2001) als kurzen und etablierten Fragebogen zur Erfassung des Konstrukts räumlicher Präsenz.

Eine wichtige methodische Thematik bei der Erhebung subjektiver Daten ist die Art der Durchführung. Fragebögen können entweder außerhalb der virtuellen Umgebung (outVRQ) – was das Abnehmen des HMD erfordert – oder direkt innerhalb der Anwendung (inVRQ) administriert werden (Alexandrovsky et al., 2020). Der Wechsel von der virtuellen in die reale Welt kann zu einem Bruch im Präsenzerleben führen, einem sogenannten Break in Presence (BIP) (Slater & Steed, 2000). Solche Unterbrechungen bergen das Risiko, die subjektiven Antworten der Teilnehmenden zu beeinflussen. Aus diesem Grund hat sich die Implementierung von Fragebögen direkt in der VR-Anwendung als relevanter Forschungsbereich etabliert. Wie sich diese Methode auf die Datenqualität auswirkt und welche konkreten Design-Ansätze existieren, wird im folgenden Kapitel [3 Aktueller Stand der Forschung](#) im Detail beleuchtet.

Grundsätzlich eignet sich VR auch für qualitative Interviews. So zeigen Mathysen und Glorieux (2021) in ihrer Studie zu VR-gestützten Interviews, dass die Befragung direkt in der virtuellen Umgebung eine valide Forschungsmethode darstellt.

Während diese Methoden essenzielle Grundlagen für die Erfassung von Erfahrungen in VR darstellen, sind sie in der Regel auf Nutzerinteraktion mit klassischen Benutzeroberflächen angewiesen. Die Datenerhebung könnte jedoch stärker in den immersiven Kontext eingebettet werden. An dieser Stelle können intelligente virtuelle Agenten zum Einsatz kommen, die Fragebögen präsentieren und durch natürliche Dialoge mit den Nutzenden neue Formen der Befragung ermöglichen, wodurch auch die Möglichkeit bestünde, soziale Aspekte zu nutzen.

2.2 Socially Interactive Agents

Als übergeordneter Sammelbegriff für soziale interaktive Agenten wird in der Literatur der Ausdruck *Soziale interaktive Agenten* (Socially Interactive Agents, SIA) verwendet (Lugrin, 2021). Darunter fallen verschiedene Konzepte und Technologien, die interaktive virtuelle oder robotische Akteure beschreiben. Die Vielzahl an Begriffen weist teilweise Überschneidungen auf, sodass eine eindeutige Abgrenzung nicht immer möglich ist. Der Schwerpunkt liegt jedoch meist entweder auf virtuellen Agenten oder auf Robotern.

Im Folgenden wird daher eine kurze Übersicht über häufig verwendete Begriffe im Bereich virtueller Agenten gegeben, angelehnt an Lugrin et al. (2021, S. 4 ff.). Da diese Begriffe in der Literatur häufig auftreten, soll die Übersicht – ohne Anspruch auf Vollständigkeit – einige Konzepte und deren Abgrenzung veranschaulichen.

- Das grundlegende Konzept eines **Agenten**, der menschliche Kommunikation versteht, seine Umgebung wahrnimmt und darin handeln kann, wurde bereits in den 1950er-Jahren entwickelt (Kay, 1984). Ein solcher Agent kann verkörpert sein, muss es aber nicht zwingend.
- Ein **Avatar** stellt die grafische Repräsentation eines Nutzers in einer virtuellen Umgebung dar. Im Unterschied zu IVAs agiert er nicht autonom, sondern wird vollständig durch den Nutzer gesteuert.
- **Virtual Humans** sind menschenähnliche künstliche Figuren mit realistischem Aussehen und Verhalten, wobei der Schwerpunkt stärker auf der visuellen Erscheinung und Simulation liegt.
- **Embodied Conversational Agents** (ECA) sind computergenerierte Charaktere, die sowohl verbal als auch nonverbal kommunizieren können. Sie treten überwiegend in virtueller Form auf, finden aber auch in der Robotik Anwendung. Der Fokus liegt hierbei stärker auf multimodaler Interaktion als auf komplexer kognitiver Intelligenz.
- Der Begriff **Intelligent Interactive Agent** (IVA) bezeichnet interaktive digitale Charaktere mit menschenähnlichen Eigenschaften, die über natürliche Modalitäten wie Sprache, Mimik und Gestik kommunizieren können. Sie sind in Echtzeit zu Wahrnehmung, Kognition, Emotion und Handlung fähig. Ein wichtiges Merkmal ist dabei ihre Autonomie und Intelligenz – im Gegensatz zu rein vorprogrammierten Reaktionsmustern (Lugrin, 2021, S. 4).
- **Socially Intelligent Agents** zeichnen sich durch menschenähnliche soziale Intelligenz und ein vertieftes Verständnis menschlicher Kognition aus. IVAs können als Teilmenge dieser Agenten mit besonderem Schwerpunkt auf virtueller Repräsentation verstanden werden.

In dieser Arbeit wird der Definition von Lugrin (2021) gefolgt und der Begriff intelligenter virtueller Agent (oder IVA) zur Bezeichnung virtueller Agenten verwendet.

Ein IVA-System besteht in der Regel aus modularen Komponenten. Dazu gehören neben grundlegenden menschlichen Kommunikationsfähigkeiten (wie Gesprächswechsel und Unterbrechungen) verschiedene Eingabemethoden (z. B. automatische Spracherkennung, Gestenerkennung, natürliche Sprachverarbeitung, Informationsextraktion, Absichtserkennung) sowie Ausgabemethoden wie visuelle 3D-Verkörperung mit Gesichtsausdrücken, Gestik und Reaktionen oder Sprachsynthese (Andreev & Cattan, 2023).

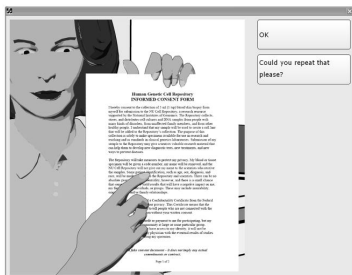
Zur Entwicklung solcher Systeme stehen verschiedene Frameworks bereit, darunter das *Virtual Human Toolkit* (Hartholt et al., 2013) oder *LiveKit Agents* (LiveKit, n. d.), die die Implementierung kompletter IVAs oder einzelner Teilkomponenten unterstützen. Alternativ können IVAs jedoch auch aus individuell ausgewählten Modulen zusammengesetzt werden, ohne auf ein spezifisches Framework zurückzugreifen.

Aufgrund der vielfältigen Einsatzmöglichkeiten werden IVAs heute in zahlreichen Bereichen genutzt. So finden sie beispielsweise als virtuelle Kundenberater, im Bildungsbereich, im Gesundheitswesen oder als Gesprächspartner Verwendung (Todericiu, 2025).

2.2.1 Darstellung und Verhalten

Die potenziellen Darstellungsformen für einen Agenten sind vielfältig und umfassen abstrakte Darstellungsformen sowie Cartoon-, Tier-, Roboter- oder menschenähnliche 2D- oder fotorealistische 3D-Modelle. Abbildung 2.5 zeigt drei Beispiele der unterschiedlichen Darstellungsweisen.

Die Darstellungs- und Verhaltensweise, aber auch die Beleuchtung eines Agenten hat Einfluss darauf, wie der Agent wahrgenommen wird, unter anderem in Bezug auf die Wahrnehmung von Persönlichkeit, Emotionen, Vertrauen und Zuversicht. Die stetige Verbesserung der Darstellungsmöglichkeiten, vor allem im Echtzeit-Rendering, sowie die Integration multimodaler Kommunikationsfähigkeiten, ermöglichen zunehmend realistischere Darstellungen



(a) Bickmore et al., 2009



(b) Ferstl et al., 2021



(c) Schmidt et al., 2024

Abbildung 2.5: Verschiedene Darstellungsweisen von Agenten: (a) eine Krankenschwester als grafischer Agent, (b) ein humanoider Roboter-Agent und (c) ein menschlicher Agent

bis hin zum Fotorealismus. Allerdings steigen mit der Zeit auch die Erwartungen an die Darstellungsqualität (McDonnell & Mutlu, 2021, 134f).

Weidner et al., 2023 empfehlen in ihrer systematischen Übersicht von 72 Studien unter anderem möglichst realistische Ganzkörperdarstellung, konsistente Bewegungen und Animationen sowie anwendungsspezifische Auswahl des Stils, um Ablenkungen zu vermeiden.

Ein wichtiger Aspekt bei der Gestaltung virtueller Agenten ist das sogenannte Uncanny Valley, ein Phänomen, das erstmals vom Robotiker Masahiro Mori in den 1970er-Jahren beschrieben wurde (Mori et al., 2012). Es beschreibt die Beobachtung, dass menschenähnliche Darstellungen bis zu einem gewissen Grad positiv wahrgenommen werden, jedoch eine plötzliche Abnahme der Akzeptanz eintritt, sobald die Darstellung fast, aber nicht vollständig menschlich wirkt (vgl. Abbildung 2.6). Die subtilen Inkonsistenzen in Mimik, Bewegung oder Erscheinung können Gefühle von Unbehagen, Fremdheit oder sogar Ekel auslösen (McDonnell & Mutlu, 2021, 130ff).

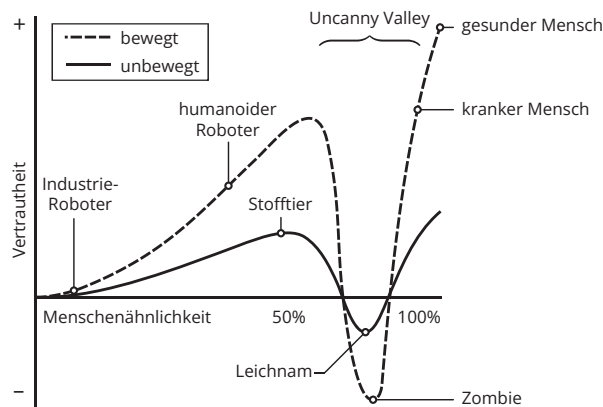


Abbildung 2.6: Darstellung des „Uncanny Valley“-Konzepts: Mit zunehmender Menschenähnlichkeit steigt die Vertrautheit künstlicher Figuren, fällt ab einem gewissen Punkt bei subtilen Unstimmigkeiten jedoch abrupt ab. (Adaptiert aus Mori et al., 2012)

2.2.2 Gestik und Mimik

Die nonverbale Kommunikation, insbesondere durch Gestik und Mimik, ist ein entscheidender Bestandteil menschlicher Interaktion und trägt maßgeblich zur Natürlichkeit und Glaubwürdigkeit von IVAs bei.

Gestik bezeichnet die spontanen, sprachbegleitenden Bewegungen, die hauptsächlich mit Händen und Armen ausgeführt werden (Saund & Marsella, 2021, S. 213). Diese Gesten erfüllen vielfältige kommunikative Funktionen: Sie können Gesprochenes unterstreichen (Schlag-Gesten bzw. Beats), Konzepte bildlich darstellen (ikonische Gesten), abstrakte Ideen visualisieren (metaphorische Gesten), auf Objekte oder Orte verweisen (deiktische Gesten)

oder sogar Wörter vollständig ersetzen (Embleme) (Saund & Marsella, 2021). Gesten dienen dabei nicht nur dem besseren Verständnis des Zuhörers – so können beispielsweise Zeigegesten in einer kollaborativen Aufgabe die kognitive Belastung des Nutzers reduzieren (Chang et al., 2022) – sondern unterstützen auch den Sprecher selbst, indem sie den Wortabruf und die Sprachflüssigkeit fördern. Ihre Wirkung hängt stark vom korrekten Timing in Bezug auf die Sprache sowie vom kulturellen Kontext ab (Saund & Marsella, 2021). Allerdings können Gesten auch eine erhöhte Aufmerksamkeitsanforderung an den Nutzer stellen (Chang et al., 2022).

Mimik ist ebenso wichtig für die soziale Interaktion. Das Gesicht dient als primäre Schnittstelle zwischen Agent und Nutzer (McDonnell & Mutlu, 2021, S. 121). Gesichtsausdrücke übermitteln Emotionen, regulieren den Gesprächsverlauf, illustrieren die gesprochenen Inhalte und können Unsicherheit signalisieren, etwa durch das Heben der Augenbrauen oder einen abgewandten Blick (Pelachaud et al., 2021; Schmidt et al., 2024). Zur systematischen Beschreibung und Generierung von Mimik wird häufig das Facial Action Coding System (FACS) herangezogen, das Gesichtsausdrücke auf Basis der zugrundeliegenden Muskelaktivitäten zerlegt (Pelachaud et al., 2021).

Sowohl Gestik als auch Mimik müssen in enger Synchronisation mit der Sprache generiert werden, um eine kohärente multimodale Kommunikation zu gewährleisten (Pelachaud et al., 2021, S. 267). Studien zeigen, dass eine Kombination aus auditiven und visuellen Signalen die Genauigkeit der wahrgenommenen Botschaft erhöht, beispielsweise bei der Einschätzung von Unsicherheit (Schmidt et al., 2024; Swerts & Kraemer, 2005). In der Praxis kommen hierfür sowohl regelbasierte Systeme zum Einsatz, die linguistische Merkmale aus Texten extrahieren, als auch datengetriebene Ansätze, die auf großen Korpora (Sammlungen von Texten oder Sprachdaten) menschlicher Interaktionen mithilfe maschinellen Lernens trainiert werden (Pelachaud et al., 2021; Saund & Marsella, 2021).

2.2.3 Soziale Präsenz und Rapport

Wie im Abschnitt 2.1.2 (Sektion [Embodiment, Soziale Präsenz und Rapport](#)) dargelegt, sind die Konzepte der sozialen Präsenz (*Social Presence*) und des Rapports wichtige Aspekte für die Qualität virtueller Interaktionen. Während Social Presence primär das Gefühl beschreibt, in einer virtuellen Umgebung mit anderen Akteuren anwesend zu sein und in Beziehung zu stehen (Harms & Biocca, 2004), geht der Begriff des Rapports darüber hinaus und erfasst die Qualität und Tiefe der sozialen Bindung, die dabei entsteht (Gratch & Lucas, 2021).

Für intelligente virtuelle Agenten ist der Aufbau von Rapport ein entscheidender Faktor, um eine glaubwürdige, angenehme und kooperative Interaktion zu ermöglichen (Gratch & Lucas, 2021). Rapport umfasst dabei Aspekte wie Vertrauen, Sympathie und die Wahrnehmung eines harmonischen Gesprächsflusses. Empirische Studien zeigen, dass IVAs durch gezielte

verbale und nonverbale Verhaltensweisen erfolgreich Rapport aufbauen können. Dies führt nicht nur zu einer positiveren Bewertung des Agenten, sondern kann auch das Vertrauen der Nutzenden stärken und ihre Bereitschaft zur Selbstoffenbarung erhöhen, insbesondere wenn sie sich weniger stark bewertet fühlen (Lucas et al., 2014).

2.2.4 Architekturen intelligenter Agenten

Die Realisierung von IVAs erfolgt mittels einer Vielzahl von Modalitäten, Techniken und Integrationsarchitekturen, wobei weiterhin an neuen Herangehensweisen für spezifische Anwendungsprobleme geforscht wird (Kopp & Hassan, 2022). Um diese Vielfalt systematisch einzuordnen, schlagen Kopp und Hassan ein konzeptuelles Modell (siehe Abb. 2.7) multimodaler Interaktion vor, das IVA-Systeme entlang dreier Verarbeitungssäulen beschreibt:

1. die **Verarbeitung** von multimodalen Eingaben,
2. die **Zuordnung** von Antworten und
3. die **Generierung** multimodaler Ausgaben.

Diese Struktur erlaubt die Beschreibung komplexer Interaktionen mithilfe funktional differenzierter Module, welche je nach System unterschiedlich stark ausgeprägt und kombiniert sein können.

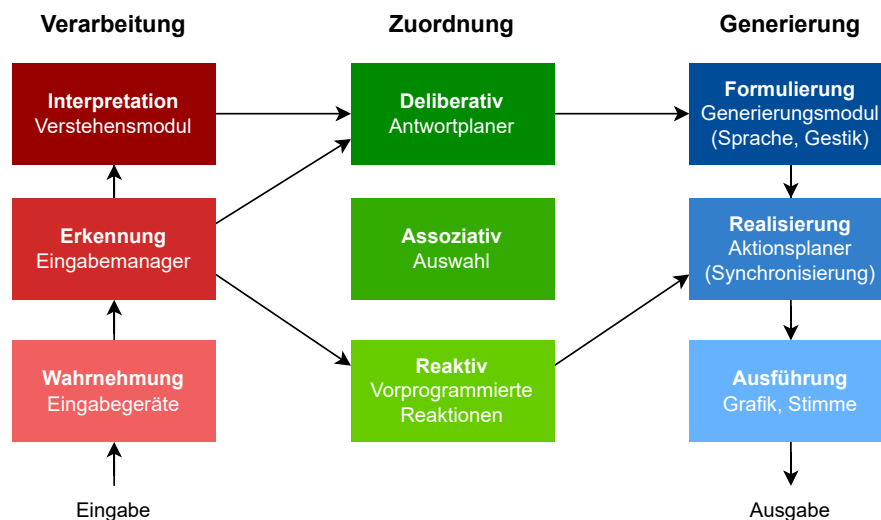


Abbildung 2.7: Verarbeitungspfade multimodalen Verhaltens in IVA-Systemen, anhand einer beispielhaften Dual-Architektur mit paralleler deliberativer und reaktiver Verarbeitung (in Anlehnung an Kopp und Hassan, 2022)

In der Praxis lassen sich gängige IVA-Architekturen, laut Kopp und Hassan (2022), grob in drei Typen unterteilen:

Single-Route-Architekturen verwenden einen linearen, assoziativen Verarbeitungsweg. Die Agenten erzeugen ihr Verhalten auf Basis geplanter sozialer Funktionen und interner Modelle, etwa zur Einschätzung von Rapport oder zur Aufgabenplanung. Der Fokus liegt auf der inhaltlichen Angemessenheit des Verhaltens, weniger auf einer dynamischen, flüssigen Gesprächsführung.

Dual-Route-Architekturen ergänzen diesen Ansatz durch zwei parallele Verarbeitungswege: einen deliberativen Pfad für komplexe Planung (die Planung möglicher neuer Reaktionen; normalerweise mit einem klassischen Natural Language Processing Modul (NLP)) und einen reaktiven Pfad für schnelle, unmittelbare Rückmeldungen auf Wahrnehmungen. So können sowohl überlegte soziale Handlungen als auch spontane Verhaltensreaktionen erzeugt werden.

Multi-direktionale, inkrementelle Architekturen gehen darüber hinaus und ermöglichen eine flexible, fortlaufende Verarbeitung. Eingaben werden stückweise (inkrementell) verarbeitet, und Informationen fließen nicht nur vorwärts, sondern auch rückwärts oder quer durch die Architektur. Dadurch können Agenten auf Unterbrechungen, Rückmeldungen und situative Veränderungen in Echtzeit reagieren.

Ein IVA-System besteht also typischerweise aus modularen Komponenten für Wahrnehmung, Verständnis, Dialogverwaltung (oder Kontextverwaltung), Entscheidungsfindung und Ausgabe. Diese Komponenten können wiederum aus mehreren Teilkomponenten bestehen (Colabianchi et al., 2023). So können zum Beispiel für die Wahrnehmung unterschiedliche Komponenten für Audio- oder Videowahrnehmung verwendet oder mit einer Kontextverwaltung zusätzliche kontextrelevante Informationen in den Dialog eingebunden werden (Maslych, Pumarada et al., 2025).

Aufbauend auf diesen allgemeinen IVA-Architekturen stellt die Sprachpipeline eine grundlegende Form der sprachbasierten Interaktion dar. Bevor sie im Detail betrachtet wird, werden im Folgenden zunächst die neuronalen Netze als technologische Grundlage moderner Sprachmodelle eingeführt.

2.3 Neuronale Netze

Neuronale Netze sind Berechnungsmodelle, deren Architektur lose von der Struktur biologischer Gehirne inspiriert ist. Sie bilden die technologische Grundlage aktueller sogenannter Künstliche Intelligenz (KI) Systeme und damit auch für die später behandelten Transformer-Architekturen und großen Sprachmodelle (LLMs) (LeCun et al., 2015). Ihre Stärke liegt darin, komplexe, nichtlineare Muster direkt aus Daten zu lernen, statt auf manuell definierte Merkmale angewiesen zu sein (LeCun et al., 2015).

Ein neuronales Netz besteht aus einem Netzwerk einfacher, miteinander verbundener Recheneinheiten, die als Neuronen bezeichnet werden. Jedes Neuron berechnet aus numerischen Eingaben eine gewichtete Summe, addiert einen sogenannten Bias-Wert und wendet eine nichtlineare Aktivierungsfunktion an (z. B. Sigmoid oder die heute verbreitete ReLU) (Jurafsky & Martin, 2025, S. 132 ff.).

Leistungsfähig werden Netze durch die Anordnung vieler solcher Neuronen in Schichten (*Layer*). In Feedforward-Netzen wird die Berechnung schrittweise von einer Eingabeschicht über eine oder mehrere verborgene Schichten (*Hidden Layers*) zur Ausgabeschicht propagiert (Jurafsky & Martin, 2025, S. 132, 138). Dabei lernen die ersten Schichten einfache Merkmale (z. B. Kanten in Bildern oder einfache Phonem-Kombinationen in Sprache), tiefere Schichten setzen diese Merkmale zu abstrakteren Repräsentationen zusammen (LeCun et al., 2015). Ein System mit vielen solchen Schichten wird als *tiefes* neuronales Netz (*Deep Neural Network*) bezeichnet (Jurafsky & Martin, 2025, S. 132).

Der Lernprozess eines neuronalen Netzes erfolgt mit einem *Backpropagation*-Algorithmus und (stochastischem) Gradientenabstieg. Dabei wird der Fehler zwischen der vom Netz vorhergesagten und der tatsächlichen Ausgabe berechnet. Dieser Fehler wird rückwärts durch das Netz propagiert und die Gewichte werden so angepasst, dass der Fehler minimiert wird (Jurafsky & Martin, 2025; LeCun et al., 2015).

Vor der Entwicklung der Transformer-Architektur dominierten verschiedene neuronale Architekturen zur Verarbeitung sequenzieller Daten:

Convolutional Neural Networks (CNNs): Diese zeigten bahnbrechende Erfolge in der Bildverarbeitung. Sie sind darauf spezialisiert, hierarchische räumliche Muster durch Faltungsoperationen (Convolutions) zu erkennen (Krizhevsky et al., 2017; LeCun et al., 2015).

Recurrent Neural Networks (RNNs): RNNs waren lange der Standard für sequenzielle Daten wie Text und Sprache. Sie verarbeiten Sequenzen schrittweise und nutzen einen internen Zustand, um Informationen aus früheren Schritten für die aktuelle Verarbeitung zu nutzen (LeCun et al., 2015). Abbildung 2.8 zeigt schematisch, wie ein RNN

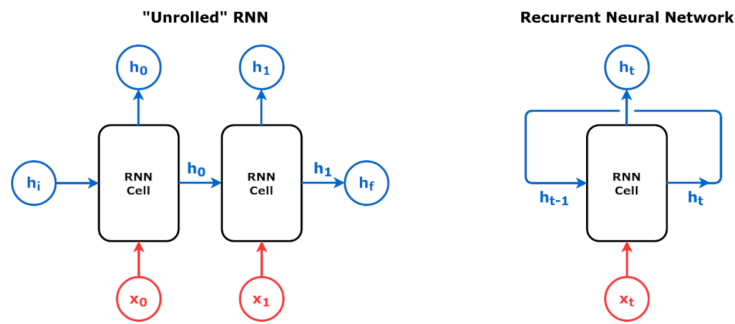


Abbildung 2.8: Schematische Darstellung eines Recurrent Neural Network (RNN). Links: entfaltete zeitliche Darstellung mehrerer Zellen, rechts: kompakte Form mit rekursiver Rückkopplung (Quelle: Godoy, 2021b, CC BY 4.0)

Informationen aus früheren Eingaben beibehält und diese bei der Verarbeitung neuer Eingaben berücksichtigt.

Long Short-Term Memory (LSTM): Eine Weiterentwicklung der RNNs, die durch spezielle Speicherzellen das Problem verschwindender Gradienten überwinden und dadurch langfristige Abhängigkeiten in Sequenzen verarbeiten können (Hochreiter & Schmidhuber, 1997; LeCun et al., 2015).

Obwohl RNNs einen wichtigen Fortschritt darstellten, stießen sie bei der Verarbeitung sehr langer Sequenzen an ihre Grenzen, da die Weitergabe von Informationen über weite Distanzen oft zu einem Verblässen des Signals führte (*Vanishing Gradient Problem*) (Jurafsky & Martin, 2025, S. 135). Genau hier setzte die Transformer-Architektur an und schuf eine neue, leistungsfähigere Grundlage für die Verarbeitung von Sequenzen.

2.3.1 Die Transformer-Architektur

Die Transformer-Architektur, 2017 von Vaswani et al. (2017) vorgestellt, revolutionierte die Verarbeitung sequenzieller Daten, insbesondere im Bereich der natürlichen Sprachverarbeitung. Ihr entscheidender Fortschritt war die Abkehr von den bis dahin dominanten rekurrenten (RNN) und sequenziellen Architekturen (wie LSTM). Diese verarbeiteten Text Token für Token (vgl. Abschnitt [Tokens](#) unten), was die Parallelisierung erschwerte und das Erfassen weitreichender Abhängigkeiten in langen Texten limitierte (Vaswani et al., 2017; Zhao et al., 2025). Abbildung 2.9 zeigt eine schematische Darstellung der Transformer-Architektur mit Encoder- und Decoder-Komponenten. Das Diagramm folgt der sogenannten pre-Layer-Normalization-Variante, die sich leicht von der ursprünglichen Implementierung von Vaswani et al. (2017) unterscheidet.

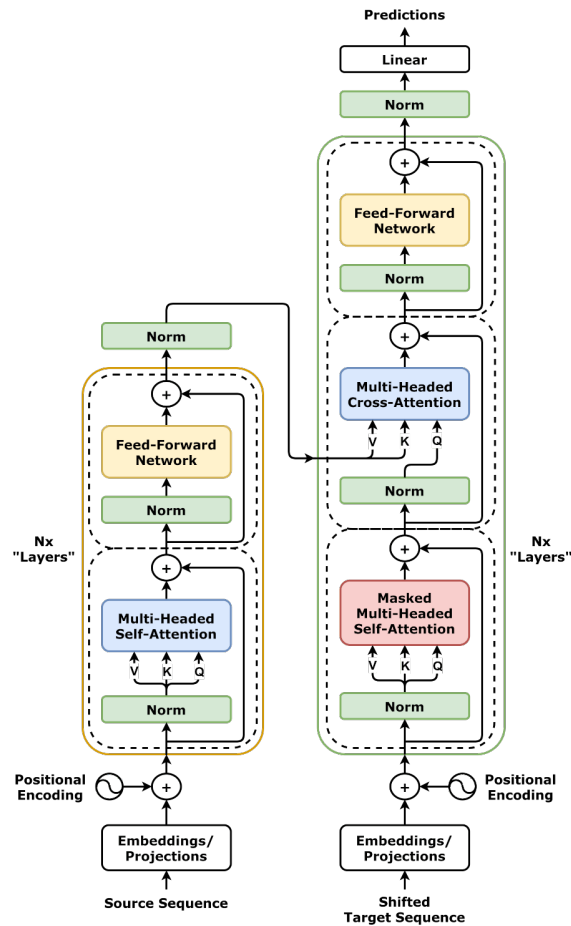


Abbildung 2.9: Schematische Darstellung einer vollständigen Transformer-Architektur mit Encoder (links) und Decoder (rechts) (Quelle: Godoy, 2021a, CC BY 4.0)

Der Transformer überwindet diese Limitierung durch den Mechanismus der Selbstaufmerksamkeit (*Self-Attention*) (Vaswani et al., 2017). Anstatt Informationen sequenziell weiterzureichen, erlaubt es die Attention dem Modell, bei der Verarbeitung eines jeden Tokens die Bedeutung aller anderen Tokens in der Eingabe abzuwägen und zu berücksichtigen. Dabei werden für jedes Token die drei Vektoren *Query* (Abfragevektor), *Key* (Schlüsselvektor) und *Value* (Wertevektor) berechnet. Die Ähnlichkeit zwischen dem Query eines Tokens und den Keys aller anderen Tokens bestimmt die Attention-Gewichte. Mit diesen Gewichten wird eine gewichtete Summe der Values gebildet, wodurch eine kontextreiche Repräsentation für jedes Token entsteht (Vaswani et al., 2017). Formal lässt sich dieser Prozess als *Scaled Dot-Product Attention* darstellen, wobei d_k die Dimension der Key- und Query-Vektoren bezeichnet:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Tokens

Damit Sprach- und Textdaten von neuronalen Netzen verarbeitet werden können, müssen sie zunächst in diskrete Einheiten zerlegt werden. Diese Einheiten werden als *Tokens* bezeichnet. Tokens können ganze Wörter umfassen (z. B. *virtuell*), bestehen in der Praxis jedoch meist aus kleineren Einheiten wie Subwörtern (z. B. *virt-uell*) oder sogar einzelnen Zeichen (Jurafsky & Martin, 2025).

Insbesondere die Zerlegung in Teilwörter ermöglicht es Modellen, mit einer festen Vokabulargröße auch seltene oder bisher unbekannte Wörter sowie komplexe Morphologie effizient zu handhaben. Nach der Tokenisierung werden die Token in Vektorrepräsentationen (*Embedding*) überführt, welche als Input für das neuronale Netz dienen. Entsprechend werden die vom Modell vorhergesagten Tokens am Ausgang wieder zu Wörtern oder Sätzen zusammengesetzt (Jurafsky & Martin, 2025). Tokens bilden damit die grundlegende Schnittstelle zwischen Rohtext und neuronaler Sprachverarbeitung.

Weitere häufige Merkmale der Transformer-Architektur sind:

Multi-Head Attention: Anstatt die Attention nur einmal auszuführen, wird der Prozess parallel in mehreren sogenannten *Heads* wiederholt. Jeder Head kann sich auf unterschiedliche sprachliche Aspekte (z. B. syntaktische oder semantische Beziehungen) fokussieren, was zu einem umfangreicheren Kontextverständnis führt (Vaswani et al., 2017).

Positionale Enkodierung: Da der Self-Attention-Mechanismus die Reihenfolge der Tokens ignoriert, werden dem Input zusätzliche Positionsinformationen hinzugefügt, damit das Modell Wortstellung und Reihenfolge innerhalb der Sequenz berücksichtigen kann (Vaswani et al., 2017).

Encoder-Decoder-Struktur: Das ursprüngliche Transformer-Modell besteht aus einem Encoder, der auf Basis dieser Repräsentationen die Eingabesequenz verarbeitet, und einem Decoder, der die Ausgabesequenz generiert (siehe Abb. 2.9, Vaswani et al., 2017). Um eine autoregressive Generierung zu gewährleisten, verwendet der Decoder eine Maskierung, die verhindert, dass ein Token beim Berechnen seiner Repräsentation auf nachfolgende Tokens zugreift.

Decoder-Only-Struktur: Viele moderne LLMs, darunter die GPT-Reihe (*Generative Pre-trained Transformer*) von OpenAI (z. B. ChatGPT), verwenden jedoch eine reine Decoder-Architektur (Zhao et al., 2025). In dieser entfällt der Encoder vollständig. Ein- und Ausgabesequenzen werden ausschließlich im Decoder verarbeitet. Dabei sorgt eine unidirektionale Attention-Maske dafür, dass jedes Token nur auf vorhergehende Tokens und sich selbst zugreifen kann (Zhao et al., 2025).

Die hohe Parallelisierbarkeit und die Fähigkeit, komplexe Abhängigkeiten in Daten zu erfassen, haben dazu geführt, dass die Transformer-Architektur zur Grundlage für die leistungsfähigsten Modelle in der Sprachverarbeitung und darüber hinaus wurde. Sie bildet die Grundlage für die Entwicklung großer Sprachmodelle (LLMs), die im folgenden Abschnitt beschrieben werden.

2.3.2 Large Language Models

Large Language Models (LLMs, *große Sprachmodelle*) sind das Resultat der Skalierung der Transformer-Architektur auf eine immense Größe, oft mit hunderten Milliarden von Parametern (Zhao et al., 2025). Sie werden mit gewaltigen Mengen an Textdaten trainiert, auf deren Grundlage sie lernen, Muster in der Sprache zu erkennen. Ihr fundamentales Trainingsziel ist die Vorhersage des wahrscheinlichsten nächsten Tokens für eine gegebene Eingabe (Zhao et al., 2025, S. 26).

Dieses scheinbar einfache Prinzip der Wahrscheinlichkeitsvorhersage reicht aus, um ein breites Spektrum sprachbezogener Aufgaben zu lösen, darunter maschinelle Übersetzungen, Textzusammenfassungen, die Beantwortung von Fragen oder das Führen von Dialogen. Durch das Training auf umfangreichen Datensätzen verdichten die Modelle statistische Regularitäten und Zusammenhänge in ihren Gewichten. Dadurch können sie in der Praxis, in gewissem Maße, als eine Art komprimiertes Abbild des in den Trainingsdaten enthaltenen Wissens betrachtet werden (Zhao et al., 2025).

Ein entscheidendes Merkmal von LLMs ist, dass eine Vergrößerung der Modell- und Datenskalierung nicht nur zu einer quantitativen Leistungssteigerung führt, sondern auch qualitative Sprünge bewirkt. Ab einer gewissen Größe zeigen LLMs sogenannte *emergente Fähigkeiten*, die in kleineren Modellen nicht oder nur in Ansätzen vorhanden sind (Zhao et al., 2025). Dieses Phänomen wird in der Literatur als „Scaling Law“ beschrieben und gilt als wesentlicher Faktor des aktuellen Erfolgs von LLMs.

Zu den wichtigsten emergenten Fähigkeiten gehören das *In-Context Learning*, bei dem das Modell aus Beispielen, die in der Eingabe (*Prompt*) stehen, neue Aufgabenstrukturen ableiten kann, sowie das *Instruction Following*, also die Fähigkeit, komplexe Anweisungen präzise zu befolgen, ohne dass diese explizit in den Trainingsdaten enthalten sein müssen (Zhao et al., 2025).

Da das Ziel der Tokenvorhersage allein jedoch nicht sicherstellt, dass die generierten Inhalte stets den menschlichen Absichten entsprechen, können LLMs auch unerwünschte Ausgaben wie sachliche Fehler, Falschinformationen oder toxische Sprache produzieren (Ouyang et al., 2022; Zhao et al., 2025). Um dem entgegenzuwirken, werden die Modelle in einem zweiten Schritt durch *Instruction Tuning* und *Reinforcement Learning from Human Feedback* weitertrainiert (Ouyang et al., 2022). Dabei bewerten Personen Modellantworten, sodass das

Modell lernt, seine Ausgaben stärker an den Präferenzen von Nutzer:innen auszurichten und hilfreicher, sicherer und kontextangemessener zu agieren.

Temperatur

Die sogenannte *Temperatur* ist ein Parameter, der die Zufälligkeit der Modellausgabe steuert, indem er die Wahrscheinlichkeitsverteilung für das nächste Token anpasst (Li et al., 2025; Renze & Guven, 2024). Kleinere Temperaturwerte machen die Ausgabe deterministischer und fokussierter, was für Aufgaben ideal ist, die Präzision und Fakten erfordern. Größere Temperaturen hingegen erhöhen die Kreativität und Vielfalt der Ergebnisse, steigern aber auch die Wahrscheinlichkeit für sachlich falsche oder unsinnige Antworten, sogenannte Halluzinationen (Renze & Guven, 2024). Der optimale Wert ist stark von der jeweiligen Aufgabe und dem verwendeten Modell abhängig, wobei größere Modelle oft unempfindlicher auf Temperaturänderungen reagieren (Li et al., 2025).

Halluzinationen

Unter dem Begriff Halluzination (*Hallucination*) wird in der KI-Forschung die Generierung von Inhalten verstanden, die unsinnig sind oder nicht dem bereitgestellten Quellinhalt entsprechen (Ji et al., 2023). Dieses Phänomen ist primär aus der Textgenerierung durch LLMs bekannt, kann jedoch auch in anderen Systemen wie der automatischen Spracherkennung auftreten. Koenecke et al. (2024) zeigen beispielsweise, dass im getesteten ASR-Modell Whisper von OpenAI (vgl. Abschnitt 5.4.3 [Automatic Speech Recognition](#)) in rund 1,4 % der Fälle vollständige Sätze erzeugt wurden, die im Audio nicht vorkamen und teilweise problematische Inhalte enthielten.

Die bisherigen Ausführungen behandelten die grundlegenden Konzepte aktueller KI-Modelle für die Sprachverarbeitung. Im folgenden Abschnitt wird mit der Sprachpipeline eine Architektur vorgestellt, die die Verarbeitung gesprochener Sprache in modularen Schritten beschreibt.

2.4 Sprachpipeline

Eine Sprache-zu-Sprache-Pipeline ist ein modulares System, das gesprochene Eingaben über meist drei aufeinanderfolgende Komponenten in eine gesprochene Ausgabe umwandelt: (1) Automatic Speech Recognition (ASR), (2) ein Language Model (LM) beziehungsweise ein Natural Language Processing (NLP)-Modell und (3) eine Text-to-Speech-Synthese (TTS). In diesem Grundkonzept übernimmt jede Komponente eine klar abgegrenzte Funktion, deren Resultat jeweils an die nächste Stufe weitergegeben wird. Ein Vorteil dieses Vorgehens besteht darin, dass einzelne Komponenten unabhängig voneinander optimiert und bei Bedarf flexibel ausgetauscht oder erweitert werden können (Sriram, 2025). Abbildung 2.10 veranschaulicht das Prinzip einer solchen Sprachpipeline.

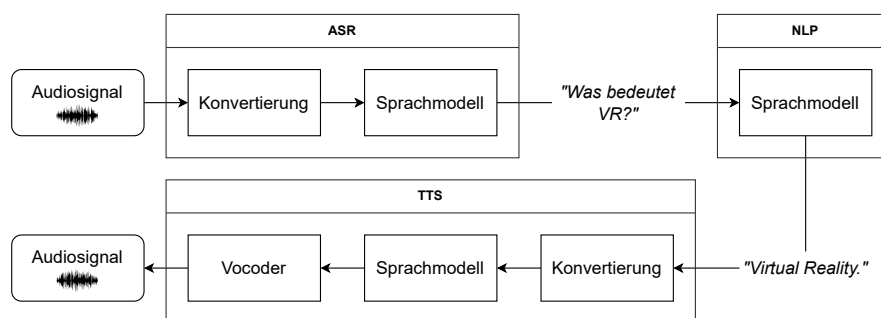


Abbildung 2.10: Schematische Darstellung der konventionellen Sprachpipeline: Transkribierung eingehender Audiosignale per ASR, Verarbeitung durch LM/NLP und Ausgabe via TTS als Audiosignal (in Anlehnung an Cui et al., 2025)

Jedoch sind mit diesem konventionellen Ansatz auch gewisse Nachteile verbunden. Es findet ausschließlich die Verarbeitung von semantischen Informationen statt, wodurch paralinguistische Informationen, also Elemente der Kommunikation, die nicht Teil der eigentlichen Sprache sind (z. B. Tonhöhe, Klangfarbe, Sprechtempo), verloren gehen. Ferner besteht das Risiko einer Fehlerfortpflanzung, falls etwa ein Fehler in der Transkription die Sprachgenerierung des NLPs beeinflusst (Cui et al., 2025).

Als Alternative werden auch End-to-End-Ansätze zu diesem modularen System erforscht. Hierbei handelt es sich um sogenannte Speech-Language-Models (SpeechLMs), die Sprache direkt in diskrete Tokens (vgl. Abschnitt 2.3.1 Tokens) codieren. Somit werden sowohl semantische als auch paralinguistische Informationen bewahrt. Anschließend werden diese Tokens autoregressiv in einem einzigen System verarbeitet, wodurch prosodische Merkmale wie Intonation und Sprechtempo direkt einbezogen werden können. Abschließend werden die vom LM generierten Tokens wieder in ein Audiosignal dekodiert, was insgesamt zu ausdrucksstärkeren und natürlicher klingenden Sprachausgaben führen kann (Cui et al., 2025).

2.4.1 Komponenten einer Sprachpipeline

Eine Sprachpipeline beginnt mit einem gesprochenen Sprachsignal, die typischerweise über ein Mikrofon erfasst wird. Das akustische Signal wird anschließend in mehreren Stufen verarbeitet, die jeweils eine klar abgegrenzte Funktion erfüllen: von der Erkennung der gesprochenen Sprache über die Interpretation bis hin zur synthetischen Ausgabe einer Antwort.

Automatic Speech Recognition

Eine Sprachpipeline beginnt mit der automatischen Spracherkennung. Dabei handelt es sich um ein seit langem bestehendes Teilgebiet der Computerlinguistik, das bereits ab den späten 1940er Jahren die ersten rudimentären Systeme hervorbrachte. In den folgenden Jahrzehnten sorgten mehrere Paradigmenwechsel für erhebliche Fortschritte, sodass ASR heute zwar nicht als vollständig gelöst gilt, aber in vielen praktischen Anwendungen zuverlässig funktioniert (Jurafsky & Martin, 2025).

Für die Spracherkennung wird das eingehende Audiosignal zunächst vorverarbeitet, indem es abgetastet, gefiltert und transformiert wird. Anschließend werden die Energieanteile berechnet und in Merkmalsvektoren überführt. Diese dienen als Grundlage für ein Sprachmodell, das die Sequenzen schrittweise in die wahrscheinlichsten Wörter und Sätze übersetzt. Das Ergebnis dieses Prozesses ist ein Texttranskript, das die erkannte Sprache in schriftlicher Form wiedergibt.

Moderne ASR-Systeme basieren fast ausschließlich auf tiefen neuronalen Netzen, die überwiegend als End-to-End-Modelle realisiert sind. Dabei kommen heute meist Transformer-Architekturen in Form von Encoder-Decoder-Systemen zum Einsatz (Jurafsky & Martin, 2025, S. 354 ff.). Zwei grundlegende Paradigmen sind Encoder-Decoder-Architekturen mit *Attention* und Modelle, die mit einer Connectionist Temporal Classification (CTC)-Funktion trainiert werden. Encoder-Decoder-Systeme (vgl. Abschnitt 2.3.1 Die Transformer-Architektur) erzielen meist eine höhere Genauigkeit, da sie den gesamten Eingabekontext berücksichtigen. Auf CTC basierende Modelle hingegen kommen ohne explizite Synchronisierung zwischen Eingabe und Zielsequenz aus. Dadurch eignen sie sich besonders für Streaming-Anwendungen, bei denen fortlaufend während der Eingabe bereits Ausgaben erzeugt werden (Graves et al., 2006; Jurafsky & Martin, 2025).

Die für das Training der Modelle genutzte Datenmenge variiert dabei erheblich, von wenigen Stunden annotiertem Audiomaterial bis zu hunderttausenden oder Millionen Stunden annotiertem oder nicht annotiertem Audiomaterial (Radford et al., 2022). Die Qualität eines ASR-Systems wird üblicherweise anhand der Word Error Rate (WER) gemessen, welche die Levenshtein-Distanz (Editierdistanz) zwischen Systemausgabe und Referenztranskription

angibt (Jurafsky & Martin, 2025). Dabei handelt es sich um ein rein quantitatives Maß, das lediglich den Anteil fehlerhafter Wörter erfasst, jedoch nicht die Art oder den Inhalt der Fehler (Koenecke et al., 2024) (vgl. Abschnitt 2.3.2 [Halluzinationen](#)).

Die automatische Spracherkennung liefert als Ausgabe eine textuelle Repräsentation des ursprünglich akustischen Signals. Dieser Text bildet die Grundlage für die nächste Verarbeitungsstufe: die Analyse und Weiterverarbeitung durch Methoden der natürlichen Sprachverarbeitung.

Natural Language Processing

Der transkribierte Text wird anschließend von einer Komponente des Natural-Language-Processings (auf Deutsch meist Computerlinguistik oder linguistische Datenverarbeitung genannt) verarbeitet. NLP kann in drei Aufgabengebiete eingeteilt werden: Sprachverstehen (das inhaltliche Erfassen der Nutzereingabe), Sprachgenerierung (das Formulieren einer passenden Antwort) und Dialogmanagement (die Steuerung des Gesprächsverlaufs zwischen zwei oder mehreren Akteuren) (Hartholt & Mozgai, 2022, S. 21).

Viele typische Aufgaben des NLP – wie etwa die Beantwortung von Fragen, Textzusammenfassung, Sentimentanalyse oder maschinelle Übersetzung – lassen sich als Wortvorhersage formulieren, weswegen in modernen Systemen häufig ein LLM (siehe Abschnitt 2.3.2 [Large Language Models](#)) verwendet wird (Jurafsky & Martin, 2025, S. 220). Das Modell interpretiert die Eingabe und erzeugt auf Grundlage seines Trainings sowie der aktuellen Aufgabenstellung den wahrscheinlichsten Ausgabertext (Cui et al., 2025). Moderne LLMs – insbesondere in Form dialogorientierter Systeme wie Chatbots – können dabei alle drei zuvor genannten Aufgabenbereiche des NLP (Sprachverstehen, Sprachgenerierung und Dialogmanagement) in einer integrierten Architektur übernehmen.

In einer Sprache-zu-Sprache-Pipeline wird der vom NLP generierte Text für die nächste Verarbeitungsstufe an eine Sprachsynthese weitergegeben.

Text-to-Speech

Der vom NLP erzeugte Ausgabertext wird in einem Text-zu-Sprache-Synthese-System (Text-to-Speech (TTS)) in gesprochenes Audio umgewandelt. Sprachsynthese zählt zu den ältesten Forschungsfeldern der Sprachverarbeitung und reicht bis zu physikalischen Modellen im 18. Jahrhundert zurück. Während frühe Verfahren auf klassische Syntheseansätze wie die Formantsynthese setzten, die auf den Resonanzfrequenzen der Sprachorgane basieren, nutzen heutige Systeme fast ausschließlich neuronale Encoder-Decoder-Architekturen. Meist werden

rekurrente Netze (LSTM) oder zunehmend Transformer-Modelle eingesetzt, um natürliche und flüssige Sprache mit hoher Qualität zu erzeugen (Jurafsky & Martin, 2025, S. 356).

Ein TTS-System funktioniert in ähnlicher Weise wie ein umgekehrtes ASR-System: Der Encoder wandelt den Text in eine abstrakte Codierung um, aus der der Decoder ein Frequenzspektrum erzeugt. Dieses wird anschließend von einem neuronalen Vocoder in eine Wellenform umgesetzt. Zuvor erfolgt eine Normalisierung, die Zahlen, Abkürzungen oder Symbole in sprechbare Formen überführt (Jurafsky & Martin, 2025).

Die zugrunde liegenden Trainingsdaten bestehen typischerweise aus stundenlangen Sprachaufnahmen einzelner Sprecher:innen, wobei moderne Systeme zunehmend Multi-Speaker-Daten nutzen, um mehrere Stimmen nachahmen oder flexibel zwischen ihnen wechseln zu können (Jurafsky & Martin, 2025). Darüber hinaus können fortschrittliche TTS-Modelle Prosodie, Stimmcharakteristika und sogar emotionale Ausdrucksweisen anpassen, um die Sprachausgabe natürlicher und kontextgerechter wirken zu lassen (Sriram, 2025).

Ein Forschungsziel ist die Reduktion von Latenz, sodass Sprachsynthese in Echtzeit für interaktive Anwendungen möglich wird. Die Qualität von TTS-Systemen wird in der Regel manuell von Menschen bewertet, da automatisierte Verfahren bislang nicht geeignet sind (Jurafsky & Martin, 2025).

Damit bildet die Sprachsynthese die abschließende Komponente der Sprachpipeline: Ausgehend von einem akustischen Nutzereingang wurde Sprache erkannt, verarbeitet und schließlich wieder als gesprochene Ausgabe bereitgestellt.

2.4.2 Weitere Aspekte aus der Computerlinguistik

Neben den oben genannten Hauptkomponenten der Sprachverarbeitung gibt es noch eine Reihe weiterer Aspekte einer Sprachpipeline, die je nach Anwendungsfall integriert werden können. Im Folgenden werden einige dieser ergänzenden Aspekte vorgestellt.

Wake-Word-Detection: Wake-Word-Detection ist ein Verfahren, das die Aktivierung von sprachgesteuerten Systemen wie Sprachassistenten durch ein bestimmtes Schlüsselwort ermöglicht. Um den Datenschutz zu gewährleisten, erfolgt die Erkennung lokal auf kleinen, ressourcenschwachen Geräten. Solche Systeme nutzen meist klassische ASR-Merkmalsextraktion und ein kompaktes Klassifikationsmodell (Jurafsky & Martin, 2025).

Voice Activity Detection: Voice Activity Detection (VAD) erkennt, ob ein kurzes Zeitfenster eines Audiosignals Sprachanteile enthält oder nicht. Einfache Systeme basieren auf Energieschwellen, wodurch sie in lauter Umgebung anfällig für Fehlklassifikationen

sind. Aktuelle Ansätze nutzen neuronale Modelle, um auch bei Störgeräuschen verlässlich Sprachsegmente zu erkennen und natürliche Sprecherwechsel zu ermöglichen (Castillo-López et al., 2025; Raux, 2008).

Turn-Taking: Als Turn-Taking wird der geregelte Wechsel von Sprecher- und Zuhörerrollen in Gesprächen bezeichnet. In menschlicher Kommunikation erfolgt der Sprecherwechsel in der Regel innerhalb von ca. 200 ms, aktuelle Sprachsysteme reagieren jedoch wesentlich langsamer. Moderne Modelle nutzen daher multimodale Eingaben wie Betonung, Timing oder Blickrichtung, um natürliche Übergänge zu ermöglichen und Überlappungen oder lange Pausen zu vermeiden.

Neben der Vorhersage von Sprecherwechseln umfasst Turn-Taking auch subtile Reaktionen wie **Backchannels** (z. B. „hmm“ oder „okay“), bei denen der Zuhörer Feedback gibt, ohne das Gespräch zu übernehmen. Damit Rückkanäle nicht fälschlicherweise als Sprecherwechsel interpretiert werden, müssen sie vom System erkannt und korrekt eingeordnet werden (Castillo-López et al., 2025).

Speaker Recognition: Die Sprechererkennung dient der Identifikation einer Stimme. Sie unterteilt sich in die Kategorien **Sprecherverifikation** (*Spricht Person X oder nicht?*) und **Sprecheridentifikation** (*Welche von n Personen spricht?*) (Jurafsky & Martin, 2025).

Speaker Diarization: Die Sprecherdiarisierung bestimmt in Mehrpersonengesprächen, wann gesprochen hat. Typischerweise kommen VAD, Sprecher-Embedding und Clusteranalyse zum Einsatz. Neuere Ansätze nutzen dagegen End-to-End-Modelle zur direkten Zuweisung von Sprecherkennzeichnungen (Jurafsky & Martin, 2025).

Language Identification: Bei der Sprachidentifikation wird die gesprochene Sprache eines Audiosignals erkannt (Jurafsky & Martin, 2025).

Die beschriebenen Konzepte und Technologien bilden die theoretische Grundlage für sprachbasierte Interaktionen mit intelligenten virtuellen Agenten. Darauf aufbauend wird im nächsten Kapitel der aktuelle Stand der Forschung beleuchtet, um bisherige Ansätze und offene Herausforderungen im Bereich der Fragebogenerhebung in Virtual Reality einzuordnen.

3 Aktueller Stand der Forschung

Die vorliegende Arbeit positioniert sich im Schnittpunkt von VR, KI und empirischer Datenerhebung. Um den Beitrag dieser Forschung einordnen zu können, wird im Folgenden der aktuelle Stand der Wissenschaft in den relevanten Teilbereichen dargestellt.

Zunächst wird die Herausforderung der Fragebogenerhebung in immersiven Umgebungen thematisiert. Darauf aufbauend wird die Sprachinteraktion mit virtuellen Agenten als Lösungsansatz vorgestellt. Anschließend werden die spezifischen technischen und interaktiven Herausforderungen bei der Implementierung von Systemen mit modernen Sprachmodellen (LLMs) erörtert. Eine abschließende Synthese fasst den Forschungsstand zusammen und leitet die Forschungslücke für diese Arbeit her.

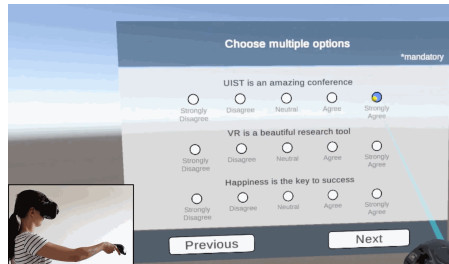
3.1 Fragebogenerhebungen in VR

Die Durchführung von Befragungen direkt in der virtuellen Umgebung (inVRQ) hat sich als methodisch vorteilhaft erwiesen, um die Validität von VR-Studien zu erhöhen. Ein Grund dafür ist die Vermeidung sogenannter BIPs – Störungen des Immersionsgefühls, die auftreten, wenn Teilnehmende das HMD abnehmen müssen, um einen Fragebogen auf Papier oder einem Computer auszufüllen (Putze et al., 2020; Schwind et al., 2019). Solche Unterbrechungen können nicht nur zu Desorientierung und Verunsicherung führen (Knibbe et al., 2018), sondern auch die Ergebnisse von subjektiven Selbstberichten verzerren.

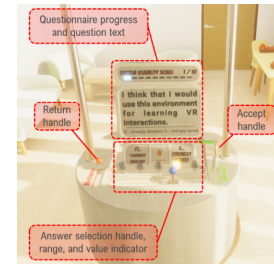
Studien zeigen, dass In-VR-Fragebögen die Konsistenz der Antworten erhöhen, ohne die Mittelwerte im Vergleich zu traditionellen Methoden zu verändern (Putze et al., 2020; Schwind et al., 2019). Physiologische Messungen zeigen zudem, dass die Intensität von BIPs messbar reduziert wird, ohne die Spielerfahrung negativ zu beeinflussen (Putze et al., 2020). Obwohl es noch keine etablierten Design-Standards gibt, werden In-VR-Befragungen von Nutzenden oft als ansprechender und natürlicher wahrgenommen als externe Webformulare, ohne dabei die Gebrauchstauglichkeit oder das Präsenzepfinden zu beeinflussen (Alexandrovsky et al., 2020; Safikhani et al., 2021). Abbildung 3.1 zeigt exemplarisch verschiedene Ansätze zur visuellen Gestaltung von In-VR-Fragebögen in der Literatur.



(a) Schwind et al., 2019



(b) Feick et al., 2020



(c) Safikhani et al., 2021

Abbildung 3.1: Beispiele für verschiedene Darstellungsweisen von inVRQs: (a) Nachbildung eines physischen Setups, (b) 2D-Interface als schwebende UI und (c) in die Umgebung integrierter 3D-Fragebogen mit interaktiven Elementen

Allerdings beschränken sich die meisten Implementierungen von inVRQs auf geschlossene Formate wie Likert-Skalen, bei denen Fragen auf einer mehrstufigen Skala beantwortet werden können (Safikhani et al., 2024). Die Erfassung von offenen Freitextantworten bleibt eine große Herausforderung, was primär an den Schwierigkeiten bei der Texteingabe in VR liegt. Das Tippen auf virtuellen Tastaturen ist im Vergleich zu physischen Pendants deutlich langsamer und kognitiv belastender (Alexandrovsky et al., 2021; Knierim et al., 2018). Die Autor:innen betonen, dass sich die meisten VR-Befragungstools aufgrund der unständlichen Texteingabe auf geschlossene Frageformate beschränken. Da offene Fragen für qualitatives Feedback jedoch unverzichtbar sind, besteht hier ein entscheidendes Forschungsanliegen (Alexandrovsky et al., 2021). Als vielversprechende Alternative wird die Nutzung von Spracheingabe diskutiert, um auch reichhaltige, qualitative Daten niedrigschwellig und nutzerfreundlich in VR zu erheben (Alexandrovsky et al., 2021; Safikhani et al., 2024).

3.2 Sprachinteraktion mit IVAs als Lösungsansatz

Spracheingabe hat sich in den letzten Jahren zu einem wertvollen und zunehmend genutzten Interaktionskanal in VR entwickelt, insbesondere in Lern- und Trainingsanwendungen. Eine systematische Übersichtsarbeit von Alghamdi und Cristea (2024) zeigt, dass Sprache vor allem in drei Bereichen eingesetzt wird: zur Kommunikation zwischen Nutzenden, zur direkten Steuerung der Umgebung und zur Interaktion mit virtuellen Agenten. Technisch dominieren dabei Cloud-basierte Speech-to-Text-Dienste, während NLP vor allem zur Intent-Erkennung und Dialogführung bei virtuellen Agenten zum Einsatz kommt. Fortschritte in LLMs ermöglichen heute offene, natürliche Sprachinteraktionen ohne ein festes Vokabular, wodurch sie zu einer wertvollen Schnittstelle für intelligente Assistenten werden (Wang et al., 2025).

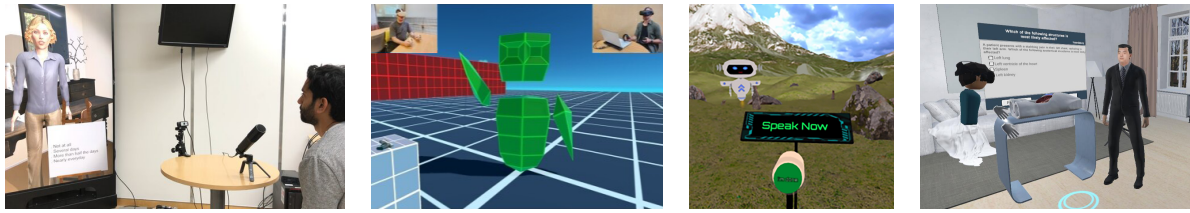
Der Einsatz von IVAs zur Administration von Fragebögen ist außerhalb von VR in den Bereichen Psychologie und Gesundheit bereits erforscht und validiert. Studien zeigen, dass die erhobenen Daten äquivalent zu traditionellen Methoden sind (Jaiswal et al., 2019) und dass Befragte, insbesondere bei sensiblen Themen, oft ehrlicher und offener antworten, wenn sie mit einem Agenten, statt mit einem Menschen interagieren (DeVault et al., 2014; Lucas et al., 2014, 2017). IVAs werden zudem als natürlicher und interaktiver wahrgenommen als klassische Online-Formulare (Soni et al., 2022). Darüber hinaus verweisen Gratch und Lucas (2021) auf mehrere Arbeiten, die darauf hindeuten, dass virtuelle Agenten mehr Offenheit und Selbstoffenbarung fördern – etwa durch längere Interaktionszeiten und umfangreichere sprachliche Äußerungen der Teilnehmenden.

Diese Erkenntnisse werden zunehmend auf den VR-Kontext übertragen. So untersuchten Young et al. (2024) die Validität von qualitativen Interviews, die direkt in VR von einem menschlichen Interviewer geführt wurden. Sie verglichen Face-to-Face-Gespräche mit zwei VR-Bedingungen: einmal ohne und einmal mit visueller Repräsentation des Interviewers durch einen Avatar (siehe Abb. 3.2b). Die Ergebnisse belegen, dass In-VR-Interviews eine valide Alternative darstellen, sofern der Interviewer visuell repräsentiert ist, um Rapport und psychologische Involviertheit zu gewährleisten. Eine wichtige Einschränkung der Studie war jedoch, dass die vorherige Bekanntschaft der Teilnehmenden mit dem menschlichen Interviewer die Ergebnisse signifikant beeinflusste: Teilnehmende, welche die interviewende Person kannten, bewerteten die Erfahrung positiver und berichteten über einen höheren Rapport sowie eine stärkere Präsenz. Die Autoren schlagen vor, dieses Problem der Voreingenommenheit durch eine bestehende soziale Beziehung mithilfe eines Stimmverzerrers zu adressieren (Young et al., 2024). In dieser Arbeit wird stattdessen ein standardisierter, neutraler IVA als mögliche Lösung vorgeschlagen.

Aufbauend auf der Machbarkeit solcher Interaktionen zeigen erste Studien das Potenzial von KI-gestützten Agenten in VR-Szenarien. Yeo et al. (2024) setzten einen KI-Agenten in VR ein, um eine immersive Form von kognitiver Verhaltenstherapie und motivierender Gesprächsführung zu ermöglichen. Die Studie belegte die hohe Machbarkeit, Akzeptanz und Offenheit der Befragten gegenüber dem therapeutischen Dialog mit dem Agenten.

In einem direkten Vergleich zwischen einem sprechenden, KI-gestützten Avatar und einer rein textbasierten Bildschirmdarstellung für eine Quizaufgabe in VR fanden Chheang et al. (2024), dass Teilnehmende bei einfachen Wissensfragen mit dem Avatar signifikant bessere Ergebnisse erzielten. Die Interaktion mit dem Avatar führte zudem zu mehr Rückfragen, besonders bei komplexeren Aufgaben. Hinsichtlich Usability und Workload gab es jedoch keine signifikanten Unterschiede; beide Systeme wurden als gut benutzbar bewertet.

Auch die visuelle Darstellung von Agenten und Avataren kann Einfluss auf die Erfahrung haben: Realistisches Aussehen und Vollkörperdarstellung können Präsenz und UX verbessern, die optimale Stilwahl bleibt jedoch aufgabenspezifisch (Weidner et al., 2023). Für gesprächs-



(a) Jaiswal et al., 2019 (b) Young et al., 2024 (c) Yeo et al., 2024 (d) Chheang et al., 2024

Abbildung 3.2: Visuelle Gestaltung von IVAs in der Literatur: menschähnliche Avatare (a) auf Monitor und (d) in VR; geometrischer Avatar (b) in VR; Roboterfigur (c) in VR

orientierte IVAs sind *Stimme* und *Bewegungsrealismus* von größerer Bedeutung als reine Optik: Natürliche Stimme und konsistente Gestik steigern Sympathie und Anthropomorphie (Vermenschlichung von Tieren und Objekten); eine hochrealistische Erscheinung bei schwacher Animation kann dagegen schaden (Ferstl et al., 2021). Mehrere Studien legen aber auch nahe, dass die Art der Darstellung von Agenten keinen Einfluss auf das subjektive Präsenzgefühl hat (Butz et al., 2022; Freiwald et al., 2021).

Insgesamt spricht die Evidenz für einen neutral-realistischen, konsistent animierten Stil mit passender Lippensynchronisation (Lipsync) und hochwertiger Sprachqualität (McDonnell & Mutlu, 2021). Abbildung 3.2 zeigt exemplarisch verschiedene visuelle Gestaltungen von IVAs in der Literatur.

Trotz dieser positiven Befunde ist die Akzeptanz sprachbasierter Systeme eng mit Fragen der Privatsphäre verknüpft. Leschanowsky et al. (2024) zeigen in ihrer systematischen Übersichtsarbeit von 100 Publikationen zur Privatsphäre in dialogorientierter KI, dass zahlreiche Studien die Themen wahrgenommene Datensicherheit, Privatsphäre und Vertrauen in Conversational-AI-Systeme untersucht haben. So analysieren etwa Ali et al. (2025), wie Nutzende die Datenerhebung, -nutzung und -speicherung in LLM-basierten Plattformen wahrnehmen, und betonen insbesondere Unsicherheiten hinsichtlich Transparenz und Kontrolle. Ergänzend beschreibt Bäckström (2025), dass Sprachsysteme ein besonderes Risiko darstellen. Bäckström hebt hervor, dass cloudbasierte Verarbeitung und ständige Audioaufnahme bei Personen die Aufmerksamkeit im Hinblick auf die Privatsphäre erhöhen und damit auch das Vertrauen der Nutzenden beeinflussen können.

3.3 Herausforderungen LLM-basierter Agenten

Die Integration moderner LLMs in VR-Systeme ermöglicht zwar hochflexible und menschenähnliche Dialoge, bringt aber auch spezifische technische und interaktive Herausforderungen mit sich. Eine Übersichtsarbeit von Özkaya et al. (2025) verdeutlicht, dass die Echtzeitfähigkeit

von LLM-gestützten VR-Anwendungen durch Hürden wie Latenz, Speicherprobleme und Skalierbarkeit limitiert wird.

Die Gestaltung der Interaktion hat erheblichen Einfluss auf die kognitive Belastung der Nutzenden. Chang et al. (2022) zeigten, dass eine Audioausgabe von einem reinen Sprach-Agenten ohne visuelle Repräsentation als abrupt empfunden wird und eine hohe kurzzeitige Aufmerksamkeitsleistung erfordert. Eine visuelle Verkörperung kann die subjektiv empfundene Arbeitslast (gemessen mit NASA-TLX) senken und das Gefühl der Kopräsenz steigern. Allerdings können nonverbale Signale, wie Gesten, auch zusätzliche Anforderungen an die Aufmerksamkeit verursachen.

Die Gefahr der kognitiven Überlastung wurde auch von Vankit et al. (2025) untersucht. In ihrer Studie zu langen Konversationen (> 20 Min.) mit IVAs in VR zeigten sie, dass die Kombination von Sprache, Untertiteln und zusätzlichen Visualisierungen das Verständnis und die Beteiligung verbesserte, aber die mentale und zeitliche Belastung signifikant erhöhte. Allerdings legt die Studie nahe, dass das Hinzufügen von Untertiteln zur gesprochenen Information die auditive Wahrnehmung unterstützen kann, ohne die Arbeitslast zu erhöhen.

Eine der größten technischen Hürden bei der Nutzung von LLM-gestützten Sprachpipelines ist die Systemlatenz – die Verzögerung zwischen dem Ende der Nutzereingabe und dem Beginn der Agentenantwort. Die Wahrnehmung dieser Latenz unterliegt etablierten Grenzwerten: Reaktionen unter einer Sekunde gelten als ideal, da sie den Gedankengang des Nutzers nicht unterbrechen (Nielsen, 1994).

Bei Konversationen mit Robotern und Agenten verschiebt sich diese Erwartung jedoch. So zeigten Shiwa et al. (2008), dass Nutzer eine leichte Verzögerung von rund einer Sekunde gegenüber einer sofortigen Antwort bevorzugen, da dies menschlichen Konversationsmustern näherkommt, während nach etwa zwei Sekunden die Zufriedenheit deutlich abnimmt.

Bereits Miller (1968) formulierte in seiner Arbeit zur Mensch-Computer-Interaktion eine psychologische Grenze von etwa zwei Sekunden, nach der Wartezeiten als störend empfunden und Denkprozesse unterbrochen werden, während Verzögerungen über vier Sekunden zu einem „Abreißen des Kommunikationsfadens“ führen. Für interaktive Systeme empfahl er daher, dass Rückmeldungen spätestens nach zwei Sekunden erfolgen sollten, um die Kontinuität des Denkens zu wahren.

Ähnliche Ergebnisse zeigte Starner (2001) in Experimenten mit tragbaren Computern: Benutzer interagierten deutlich häufiger mit sprach- oder tastengesteuerten Wearables, wenn die Systemreaktion innerhalb dieser Zwei-Sekunden-Grenze („Two-Second-Rule“) lag, während bereits geringfügig längere Verzögerungen zu Frustration und seltenerer Nutzung führten (Miller, 1968; Starner, 2001). Diese Erkenntnisse bleiben auch für moderne KI-gestützte Systeme relevant (Elfleet & Chollet, 2024).

In modernen Systemen verstärkt sich dieses Problem durch cloudbasierte Verarbeitung. LLM-gestützte Sprachagenten überschreiten diese Schwelle häufig und erreichen Antwortzeiten von über drei Sekunden (Elfleet & Chollet, 2024), was die wahrgenommene Natürlichkeit und den Dialogfluss deutlich reduziert. In einer Pilotstudie ermittelten Maslych, Pumarada et al. (2025) eine durchschnittliche Antwortzeit von 3,2 s für ihr lokal laufendes System, was von Teilnehmenden nur als mäßig responsiv bewertet wurde.

Arbeiten von Elfleet und Chollet (2024) und Shiwa et al. (2008) zeigen, dass multimodales Feedback – insbesondere Füllwörter wie „Hmm, lass mich nachdenken“ in Kombination mit Gestik oder Blickbewegungen – die wahrgenommene Antwortzeit deutlich reduziert und den Gesamteindruck des Agenten verbessert. In einer systematischen Untersuchung verglichen Maslych, Katebi et al. (2025) die in Abbildung 3.3 dargestellten Strategien und zeigten, dass natürliche Füller am wirksamsten sind, während künstliche Ladeindikatoren kaum eine Verbesserung erzielen.

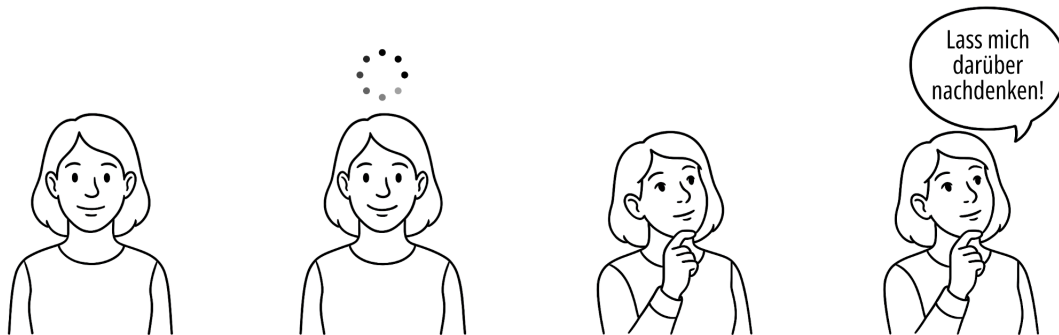


Abbildung 3.3: Gegenüberstellung von vier Strategien zum Umgang mit Systemlatenz bei virtuellen Agenten: (v. l. n. r.) keine Rückmeldung, künstlicher Ladeindikator, natürliche Darstellung durch Geste, natürlicher konversationeller Füller durch Geste und Sprache (in Anlehnung an Maslych, Katebi et al., 2025, Illustration generiert von ChatGPT, eigene Darstellung)

3.4 Synthese und Positionierung der Forschung

Der aktuelle Forschungsstand zeigt, dass die Befragung mittels sprachgesteuerter virtueller Agenten in VR eine vielversprechende Lösung für die methodischen Herausforderungen traditioneller Datenerhebungsverfahren darstellt. Während das Potenzial von LLMs zur Unterstützung vieler Schritte in der Umfrageforschung – von der Fragenerstellung bis zur Analyse – bereits erkannt wurde, identifizierten Jansen et al. (2023) die eigentliche Datenerhebung als eine wesentliche Lücke. Sie stellten fest, dass LLMs scheinbar bereits alle Aufgaben der Umfrageforschung übernehmen können, außer der Administration der Erhebung selbst:

„Seemingly, the only primary survey research task these LLMs cannot yet do is data collection, which is administering the survey instrument to the selected sample.“ (Jansen et al., 2023, S. 3). Gleichzeitig spekulieren die Autor:innen jedoch, dass auch diese Grenze bald überschritten sein könnte.

Genau an dieser Forschungslücke setzt die vorliegende Arbeit an. Durch die Entwicklung und Evaluation eines sprachbasierten, LLM-gesteuerten Interview-Agenten in VR wird untersucht, ob Sprachmodelle nicht nur unterstützend, sondern als aktive und zuverlässige Befrager eingesetzt werden können. Damit adressiert diese Studie eine wesentliche Forschungslücke und erweitert den Einsatzbereich von LLMs in der empirischen Datenerhebung um die Phase der Erhebung selbst. Sie leistet somit einen Beitrag zur methodischen Weiterentwicklung von Datenerhebungen in VR, ohne dabei die Immersion der Teilnehmenden zu beeinträchtigen, insbesondere für die Erfassung von Freitextantworten.

4 Forschungsdesign und Methodik

Das Forschungsdesign dieser Arbeit ist darauf ausgerichtet, die Möglichkeiten und Grenzen einer sprachbasierten Fragebogenerhebung in Virtual Reality mit einem intelligenten virtuellen Agenten zu untersuchen. Der Fokus liegt dabei auf der Analyse der Zuverlässigkeit, der Qualität der Antworten sowie der Nutzererfahrung in der Interaktion mit dem IVA. Um die Ergebnisse einordnen zu können, wird als Vergleichsbedingung eine klassische Befragung über eine 2D-Benutzeroberfläche in VR herangezogen. Um die Forschungsfragen zu beantworten, kommt ein experimentelles Studiendesign mit zwei unterschiedlichen, klar voneinander abgegrenzten Bedingungen zum Einsatz.

Vor der eigentlichen Studie wurde eine Pilotstudie mit wenigen Teilnehmenden durchgeführt, um technische Abläufe, Instruktionen sowie die Verständlichkeit und Vollständigkeit der Fragebögen zu überprüfen. Auf diese Weise konnten potenzielle Probleme frühzeitig identifiziert und behoben werden.

4.1 Studiendesign

Die Untersuchung folgt einem Within-Subjects-Design, bei dem alle Teilnehmenden beide Versuchsbedingungen durchlaufen:

Bedingung A: Ausfüllen der Fragebögen per sprachbasierter Interaktion mit einem intelligenten virtuellen Agenten in VR.

Bedingung B: Ausfüllen der Fragebögen per klassischer 2D-Benutzeroberfläche in VR.

Um Reihenfolge- und Lerneffekte zu minimieren, wird die Abfolge der Versuchsbedingungen systematisch ausbalanciert.

Die Stichprobengröße wurde anhand einer a-priori-Power-Analyse mit *G*Power* (Version 3.1; Faul et al., 2007) bestimmt, um eine ausreichende Teststärke für die geplanten statistischen Analysen (siehe Abschnitt 4.4.2) sicherzustellen. Bei einem mittleren erwarteten Effekt ($d_z = 0,5$), einem Signifikanzniveau von $\alpha = .05$ und einer angestrebten Teststärke von $(1 - \beta) = .80$ ergibt sich eine erforderliche Stichprobengröße von $n = 34$. Die Teilnehmenden

werden über Aushänge, persönliche Ansprache sowie interne soziale Kanäle unter Studierenden, Mitarbeitenden, Besucher:innen der Hochschule und Personen aus dem erweiterten Bekanntenkreis rekrutiert.

Die Studie wurde durch die Ethikkommission der HAW Hamburg geprüft und genehmigt. Alle Teilnehmenden erhielten vorab schriftliche Informationen über Ziel, Ablauf und Datennutzung und gaben eine informierte Einverständniserklärung ab.

4.2 Versuchsablauf

Der Studienablauf ist so gestaltet, dass eine systematische und vergleichbare Datenerhebung zwischen den beiden Versuchsbedingungen gewährleistet ist. Jede Versuchsperson durchläuft beide Bedingungen in zufälliger Reihenfolge, um Reihenfolgeeffekte zu kontrollieren. Der Ablauf gliedert sich in vier Phasen:

1. Einführung und Aufklärung:

- Begrüßung, Erklärung des Ablaufs und des VR-Szenarios, Unterzeichnung der Einverständniserklärung
- Ggf. kurze Einführung in die VR-Technik und Bedienung

2. Durchlauf pro Bedingung:

- Die Teilnehmenden absolvieren zunächst eine kurze VR-Erfahrung (ca. 3 – 4 Minuten) als Bezugsszenario.
- Anschließend füllen sie, in VR, den Fragebogen aus – je nach Bedingung über den IVA oder die 2D-UI.
- Direkt nach jeder Bedingung werden zwei zusätzliche Fragebögen außerhalb von VR am Rechner erhoben. Abgefragt werden die subjektive Arbeitsbelastung und die Gebrauchstauglichkeit der Fragebogenerhebung in VR.

3. Abschluss:

- Nach Abschluss der beiden Bedingungen folgt ein Abschlussfragebogen zu Nutzererfahrung, Präferenzen und Veränderungswünschen.
- Anschließend wird ein demografischer Fragebogen erhoben.
- Abschlussgespräch, Gelegenheit für Rückfragen, Bedankung und Verabschiedung.

4.3 Erhebungsinstrumente

Zur Datenerhebung kommen standardisierte Fragebögen, offene Fragen und technische Messungen zum Einsatz:

Nach der VR-Experience (in VR):

- Igroup Presence Questionnaire (IPQ) zur Erfassung von Präsenz
- Offene Fragen zu Bedienbarkeit und Verständlichkeit, Motion Sickness und sonstigen Bemerkungen

Nach jedem VR-Durchlauf (außerhalb von VR):

- Raw NASA Task Load Index (NASA-TLX) zur subjektiven Arbeitsbelastung, jeweils mit Bezug auf das Beantworten von Likert-Skalen und das Beantworten von Freitextfragen
- System Usability Scale (SUS) zur Gebrauchstauglichkeit des Systems

Am Ende der Studie:

- Abschlussfragebogen zu Wohlbefinden, Sicherheit, Verständlichkeit, Schwierigkeiten, Vertrauenswürdigkeit bei der Interaktion mit dem IVA; als Likert- oder Multiple-Choice-Fragen jeweils mit optionalem Freitextfeld
- Präferenzen zur Fragebogenerhebung in VR (Agent oder UI)
- Offene Abschlussfrage und Verbesserungsvorschläge
- Demografischer Fragebogen (Alter, Geschlecht, VR-/KI-Erfahrung)

Technische Messungen:

- Fehlerrate und Latenz in der Sprachpipeline
- Protokollierung (*Logging*) der Sprachpipeline-Ereignisse und Nutzerinteraktionen.

Die eingesetzten Fragebögen sind etablierte und validierte Instrumente in der VR- und HCI-Forschung:

Zur Erfassung des Präsenzepfindens in virtuellen Umgebungen wird der **Igroup Presence Questionnaire (IPQ)** (Schubert et al., 2001) eingesetzt. Er umfasst die Dimensionen räumliche Präsenz (*Spatial Presence*), Involviertheit (*Involvement*) und erlebten Realismus (*Experienced Realism*) und ist ein Standardinstrument für VR-Studien. Für die vorliegende Studie wird die deutschsprachige Version des IPQ nach Schubert et al. (2001) eingesetzt.

Die subjektive Arbeitsbelastung wird mit dem **Raw NASA Task Load Index (Raw NASA-TLX)** (Hart & Staveland, 1988) erfasst, das sechs Dimensionen berücksichtigt: mentale, physische und zeitliche Anforderungen, wahrgenommene Leistung, Anstrengung und Frustration. Es wird die Raw-Variante verwendet, bei der die sechs Dimensionen direkt auf Ratingskalen bewertet und anschließend gemittelt werden. Auf die zusätzliche Gewichtung der Dimensionen, wie sie im ursprünglichen NASA-TLX vorgesehen ist, wird verzichtet (Hart, 2006). In dieser Studie wird der NASA-TLX zweimal parallel abgefragt – jeweils mit spezifischem Bezug auf das Beantworten von Skalenfragen (Likert-Items) und das Beantworten von Freitextfragen. Dadurch kann die wahrgenommene Belastung differenziert für beide Fragetypen untersucht werden. Als Grundlage dient die offizielle Fassung der Skala von der NASA Human Systems Integration Division (NASA Human Systems Integration Division, 2022), die für diese Studie ins Deutsche übersetzt wurde.

Zur Bewertung der subjektiv empfundenen Benutzerfreundlichkeit kommt die **System Usability Scale (SUS)** (Brooke, 1996) zum Einsatz. Die SUS ist ein international etabliertes Kurzverfahren mit zehn Items, das bereits in zahlreichen Studien zur schnellen Beurteilung der Usability eingesetzt wurde. In der vorliegenden Studie wird die SUS in der Übersetzung nach Gao et al. (2020) verwendet.

Ergänzend wurden ein **Abschlussfragebogen** sowie offene Fragen entwickelt, um Eindrücke zum Wohlbefinden, zur Verständlichkeit, zu Schwierigkeiten sowie zu Präferenzen hinsichtlich der Befragungsform zu erfassen. Abschließend werden demografische Daten wie Alter, Geschlecht sowie Vorerfahrungen mit VR-, Sprachassistenten- und KI-Systemen erhoben.

Die vollständigen Fragebögen und Itemformulierungen finden sich im Anhang **B Erhebungsinstrumente (Fragebögen)**.

4.4 Hypothesen und Operationalisierung

Aufbauend auf den in Abschnitt 1.1 formulierten Forschungsfragen und dem in Kapitel 3 beschriebenen Stand der Forschung werden im Folgenden die Hypothesen dieser Arbeit hergeleitet und deren empirische Überprüfung beschrieben. Grundlage bilden Erkenntnisse zu In-VR-Fragebögen, zur Sprachinteraktion mit intelligenten virtuellen Agenten (IVA) sowie zu technischen und interaktiven Besonderheiten moderner Sprachmodelle (LLMs).

Vor diesem Hintergrund untersucht die vorliegende Arbeit, ob ein IVA eine valide und nutzerfreundliche Alternative zur Fragebogenerhebung über eine klassische 2D-UI darstellt. Daraus ergeben sich vier Hypothesenkomplexe (H1 – H4) zu den Themenbereichen Zuverlässigkeit, Datenqualität, Nutzererlebnis und Präsenzerleben.

4.4.1 Hypothesen

H1: Zuverlässigkeit

Wie zuverlässig funktioniert die Beantwortung von Fragebögen durch einen IVA in VR aus technischer und subjektiver Perspektive?

Vorarbeiten zeigen, dass IVAs valide und konsistente Datenerhebungen ermöglichen, die traditionellen Methoden entsprechen (Jaiswal et al., 2019). Befragte empfinden sie als vertrauenswürdig und reagieren offener (Lucas et al., 2014, 2017). Da moderne Sprachmodelle gesprochene Antworten zuverlässig erkennen und interpretieren können, wird erwartet, dass das im Rahmen dieser Arbeit entwickelte System zur Fragebogenerhebung technisch zuverlässig arbeitet und der IVA als verlässlicher Interaktionspartner wahrgenommen wird.

H1a: Das System erfasst die quantitativen Antworten der Teilnehmenden ohne systematische Fehler.

H1b: Die Teilnehmenden bewerten die Fähigkeit des Agenten, ihre Antworten korrekt zu verstehen, als hoch und empfinden ihn als zuverlässigen Interaktionspartner.

H2: Datenäquivalenz und Antwortqualität

Unterscheiden sich die gegebenen Antworten auf Fragebögen, wenn diese in VR durch einen IVA oder über eine klassische 2D-UI ausgefüllt werden?

Vorangegangene Studien weisen darauf hin, dass Agenten qualitativ vergleichbare oder sogar reichhaltigere Antworten erzeugen als klassische Eingabeformen (Gratch & Lucas, 2021; Lucas et al., 2017). Sprachinteraktion kann Offenheit und Spontaneität fördern und so zu inhaltlich differenzierteren Antworten führen (Yeo et al., 2024; Young et al., 2024), kann aber auch die kognitive Anforderung erhöhen (Vankit et al., 2025). Daher wird erwartet, dass die quantitativen Antworten beider Bedingungen vergleichbare Mittelwerte, in der IVA-Bedingung jedoch eine größere Streuung aufweisen. Qualitative Freitextantworten werden zudem als sprachlich umfangreicher und gehaltvoller angenommen als in der UI-Bedingung.

H2a: Die in der IVA-Bedingung erhobenen quantitativen Antwortwerte weisen eine größere Streuung und eine stärker von der Normalverteilung abweichende Verteilungsform auf als in der UI-Bedingung.

H2b: Sprachlich gegebene Freitextantworten in der IVA-Bedingung sind ausführlicher und inhaltlich reichhaltiger als über die UI eingegebene Antworten.

H3: Nutzererlebnis (UX)

Wie wird die Nutzererfahrung (UX) bei der Befragung durch einen IVA im Vergleich zur Interaktion über eine klassische 2D-UI wahrgenommen?

Sprachbasierte Interaktionen werden als natürlicher und sozialer empfunden als klassische 2D-UIs (Chheang et al., 2024; Soni et al., 2022). Während Spracheingabe die motorische Belastung reduziert, kann sie kognitiv fordernder sein (Vankit et al., 2025). Eine visuelle Verkörperung des Agenten wiederum senkt die empfundene Arbeitslast und steigert die wahrgenommene Natürlichkeit (Chang et al., 2022). Es wird daher angenommen, dass die IVA-Bedingung insgesamt eine höhere Gebrauchstauglichkeit, jedoch unterschiedliche Belastungswerte je nach Fragetyp aufweist.

H3a: Die subjektive Arbeitsbelastung bei der Beantwortung von Likert-Skalen ist in der IVA-Bedingung höher als in der UI-Bedingung.

H3b: Die subjektive Arbeitsbelastung bei der Beantwortung von Freitextfragen ist in der IVA-Bedingung niedriger als in der UI-Bedingung.

H3c: Die subjektive Gebrauchstauglichkeit der sprachbasierten Befragung in der IVA-Bedingung wird höher bewertet als die Interaktion in der UI-Bedingung.

H4: Präsenzerleben

Wie wirkt sich die Befragung durch einen IVA im Vergleich zur Interaktion über eine 2D-UI auf das subjektive Präsenzerleben in der VR-Umgebung aus?

Vorarbeiten zeigen, dass weder inVRQs noch IVAs das Präsenzempfinden signifikant verändern, sofern die Darstellung konsistent mit der VR-Umgebung bleibt (Freiwald et al., 2021; Putze et al., 2020; Schwind et al., 2019; Weidner et al., 2023). Daher wird kein signifikanter Unterschied im Präsenzerleben zwischen beiden Bedingungen erwartet.

H4: Das subjektive Präsenzerleben unterscheidet sich nicht signifikant zwischen der IVA- und der UI-Bedingung.

4.4.2 Operationalisierung

Zur Überprüfung der Hypothesen kommen standardisierte Fragebögen, offene Fragen und technische Messungen zum Einsatz. Die Studie folgt einem Within-Subjects-Design, in dem alle Teilnehmenden beide Bedingungen absolvieren (IVA vs. UI). Die verwendeten Instrumente sind etablierte Verfahren in der VR- und HCI-Forschung. Eine kompakte Beschreibung der statistischen Auswertungsverfahren folgt in Abschnitt [4.5 Statistische Methoden](#).

Technische Zuverlässigkeit (H1a): Latenzen, Transkriptionsfehler und Fehlklassifikationen der Sprachpipeline werden aus Systemprotokollen extrahiert und deskriptiv ausgewertet.

Subjektive Zuverlässigkeit (H1b): Einschätzungen zur Verständlichkeit und Vertrauenswürdigkeit des Agenten werden über Likert-Items und offene Fragen im Abschlussfragebogen erfasst.

Quantitative Datenäquivalenz (H2a): Der IPQ wird in beiden Bedingungen innerhalb der VR-Umgebung erhoben. Varianzunterschiede werden mit dem Pitman-Morgan-Test geprüft und Schiefedifferenzen per Bootstrapping analysiert.

Qualitative Antwortqualität (H2b): Drei offene Fragen innerhalb von VR werden thematisch kodiert und ausgewertet. Indikatoren sind Wortanzahl, thematische Vielfalt und Gehalt der Antworten. Diese Maße werden mittels Wilcoxon-Vorzeichen-Rang-Test geprüft.

Arbeitsbelastung (H3a/b): Zur Erfassung der empfundenen kognitiven Beanspruchung wird der Raw NASA-TLX nach jeder Bedingung, getrennt nach Fragetyp (Likert-Skalen vs. Freitextfragen), erhoben. Unterschiede werden mittels gepaarter *t*-Tests geprüft und die Signifikanzniveaus werden nach Holm-Bonferroni angepasst.

Gebrauchstauglichkeit (H3c): Die wahrgenommene Benutzerfreundlichkeit wird mit der SUS erfasst und per gepaartem *t*-Test verglichen.

Präsenzerleben (H4): Der IPQ wird in VR ausgefüllt und misst das subjektive Präsenzerleben. Die Auswertung erfolgt über gepaarte *t*-Tests.

4.5 Statistische Methoden

Zur Überprüfung der Hypothesen werden quantitative und qualitative Daten aus beiden Erhebungsbedingungen analysiert. Zunächst werden die statistischen Annahmen (z. B. Normalverteilung und Ausreißer) überprüft. Anschließend erfolgt die Hypothesenprüfung mit geeigneten Testverfahren. Alle Analysen werden auf einem Signifikanzniveau von $\alpha = ,05$ durchgeführt. Bei mehrfachen Tests wird das Signifikanzniveau nach Holm-Bonferroni (Holm, 1979) angepasst, um Fehlerkumulationen zu vermeiden. Im Folgenden werden die eingesetzten Analysemethoden kurz erläutert.

Annahmenprüfung

Normalverteilung (Shapiro-Wilk-Test)

Mit dem Shapiro-Wilk-Test (Shapiro & Wilk, 1965) wird geprüft, ob die Daten einer theoretischen Normalverteilung entsprechen. Ein signifikanter *p*-Wert ($p < ,05$) weist auf Ab-

weichungen hin, während ein nicht signifikanter Wert Normalverteilung nahelegt. Dieses Verfahren wird besonders für kleine Stichproben < 50 empfohlen.

Verteilungsform (Schiefe)

Zur Beurteilung der Form der Verteilung wird die Schiefe (γ) berechnet. Diese beschreibt, ob eine Verteilung symmetrisch ist ($\gamma = 0$) oder in eine Richtung verzerrt ($\gamma < 0$: linksschief, $\gamma > 0$: rechtsschief). Nach Bulmer (1979) können Werte unter $|\gamma| < 0,5$ als weitgehend symmetrisch, Werte zwischen $0,5 \leq |\gamma| < 1$ als moderat schief und Werte ab $|\gamma| \geq 1$ als stark schief interpretiert werden.

Interne Konsistenz (Cronbachs α)

Zur Bewertung der Zuverlässigkeit der Skalen wird Cronbachs α berechnet (Cronbach, 1951). Das Maß gibt an, wie stark die einzelnen Items einer Skala miteinander zusammenhängen und dasselbe Konstrukt erfassen. Ein höherer α -Wert weist auf eine größere Homogenität der Items und somit auf eine höhere Zuverlässigkeit der Skala hin. Nach Blanz (2021, S. 250) können folgende Angaben zur Interpretation herangezogen werden: $\alpha > ,90$ exzellent, $\alpha > ,80$ gut, $\alpha > ,70$ akzeptabel, $\alpha > ,60$ fragwürdig, $\alpha > ,50$ schlecht und $\alpha < ,50$ inakzeptabel.

Vergleichsverfahren

Parametrischer Vergleich (gepaarter t -Test)

Zur Prüfung von Mittelwertunterschieden zwischen zwei abhängigen Bedingungen wird der gepaarte t -Test eingesetzt. Das Verfahren berücksichtigt, dass dieselben Personen in beiden Bedingungen gemessen werden, und testet, ob sich die Mittelwerte signifikant unterscheiden. Ein signifikanter p -Wert ($p < ,05$) deutet auf einen systematischen Unterschied hin.

Nichtparametrischer Vergleich (Wilcoxon-Vorzeichen-Rang-Test)

Wenn die Normalverteilungsannahme verletzt ist, wird der Wilcoxon-Vorzeichen-Rang-Test (Wilcoxon, 1945) verwendet. Dieses nichtparametrische Verfahren prüft, ob die Differenzen zwischen den Messwerten symmetrisch um null verteilt sind. Ein signifikanter p -Wert ($p < ,05$) weist auf Unterschiede zwischen den Bedingungen hin.

Proportionsvergleich (Einstichproben- z -Test)

Zur Prüfung, ob beobachtete Anteile signifikant von einer Gleichverteilung abweichen (z. B. Präferenzen für IVA vs. UI), wird ein Einstichproben-Gauß-Test (Einstichproben- z -Test für Anteile) (Blanz, 2021; Eckstein et al., 1994) durchgeführt. Das Verfahren vergleicht die beobachteten Häufigkeiten mit einem theoretischen Erwartungswert von 50 % und prüft, ob die

Abweichungen zufällig oder systematisch sind. Ein signifikanter p -Wert ($p < ,05$) weist auf eine systematische Abweichung von der erwarteten Gleichverteilung hin.

Korrektur multipler Vergleiche (Holm-Bonferroni)

Werden mehrere Dimensionen eines Fragebogens getestet, erfolgt eine Alphafehler-Korrektur nach der Holm-Bonferroni-Methode (Holm, 1979), um einer Alphafehler-Kumulierung entgegenzuwirken. Dabei werden die einzelnen Tests in eine Rangfolge ihrer p -Werte gebracht und sukzessive strengere Signifikanzgrenzen angelegt, um eine Überbewertung zufälliger Effekte zu vermeiden. Dadurch wird die Wahrscheinlichkeit verringert, fälschlicherweise ein signifikantes Ergebnis zu finden (Fehler 1. Art).

Gepaarter Varianzvergleich (Pitman-Morgan-Test)

Varianzunterschiede zwischen den beiden Erhebungsformen werden mit dem Pitman-Morgan-Test geprüft (Morgan, 1939; Pitman, 1939). Das Verfahren testet die Nullhypothese gleicher Varianzen in gepaarten Messungen, indem es die Korrelation zwischen Summen- und Differenzwerten der Messpaare auswertet.

Schiefedifferenz (gepaartes Bootstrapping-Verfahren)

Unterschiede der Verteilungsform werden über die Differenz der Schiefe $\Delta\gamma = \gamma_{IVA} - \gamma_{UI}$ bewertet. Für $\Delta\gamma$ wird ein gepaartes Bootstrapping-Verfahren (Stichprobenwiederholung) mit hoher Wiederholungszahl eingesetzt. Aus der resultierenden Verteilung von $\Delta\gamma$ wird ein 95 %-Konfidenzintervall bestimmt. Ein Intervall, das 0 nicht enthält, weist auf einen systematischen Unterschied der Schiefe zwischen den Bedingungen hin.

Effektstärken

Cohens d

Für die parametrischen Tests wird die Effektstärke nach Cohen berechnet. Sie beschreibt die Größe des Mittelwertunterschieds unabhängig von der Stichprobengröße und erlaubt eine Einschätzung der praktischen Bedeutsamkeit. Die Berechnung erfolgt nach der Formel

$$d = \frac{M_1 - M_2}{SD_{diff}}$$

Nach den Konventionen von Cohen (1988) gelten Werte um $d = 0,2$ als kleine, $d = 0,5$ als mittlere und $d = 0,8$ als große Effekte.

Wilcoxon-Vorzeichen-Rang-Test Effektstärke

Für nichtparametrische Vergleiche wird die Effektstärke per Pearson Korrelationskoeffizient berechnet, die sich aus dem z-Wert des Wilcoxon-Vorzeichen-Rang-Tests nach der Formel

$$r = \frac{z}{\sqrt{N}}$$

ergibt (Fritz et al., 2012). Nach den Konventionen von Cohen (1988) gelten $|r| = 0,10$ als kleine, $|r| = 0,30$ als mittlere und $|r| = 0,50$ als große Effekte. Dieses Maß beschreibt die Stärke der beobachteten Unterschiede unabhängig von der Stichprobengröße.

4.6 Pilotstudie

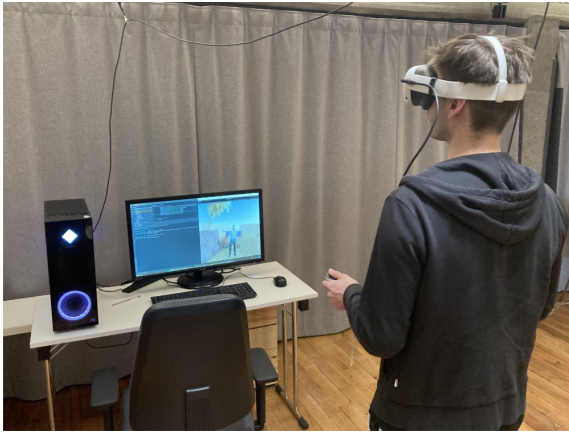
Vor der Hauptstudie wurde eine kleine Pilotstudie mit drei Teilnehmenden durchgeführt, um die technische Funktionsfähigkeit der Prototypen sowie die Verständlichkeit und Plausibilität der Erhebungsinstrumente zu überprüfen. Dabei wurden sowohl der Ablauf der VR-Erfahrung, sowie der Sprachbefragung als auch die inhaltliche Gestaltung der Fragebögen getestet. Die Datensätze der Pilotstudie wurden nicht ausgewertet und flossen nicht in die Hauptanalyse ein.

Das qualitative Feedback der Teilnehmenden lieferte wertvolle Hinweise für die Optimierung des Studiendesigns – insbesondere bezüglich des NASA-TLX, welcher in der Pilotstudie pro Bedingung (IVA und UI) nur einmal abgefragt wurde. Die Teilnehmenden äußerten, dass sie vor allem in der UI-Bedingung den NASA-TLX je nach Fragetyp (Likert-Skalen vs. Freitextfragen) unterschiedlich beantworten würden, da die wahrgenommene Belastung bei Freitextfragen deutlich höher eingeschätzt wurde. Daher wurde der NASA-TLX in der Hauptstudie für jede Bedingung zweimal erhoben – jeweils mit spezifischem Bezug auf die Beantwortung der Likert-Fragen und der Freitextfragen.

4.7 Versuchsaufbau

Der Versuchsaufbau umfasst die für die Durchführung der Studie erforderliche technische Infrastruktur sowie deren Anordnung im Testumfeld. Das Gesamtsystem besteht aus einer Sprachpipeline, die die Sprachverarbeitung, Protokollierung und Steuerung der Fragebögen übernimmt, und einem VR-Szenario, in dem die visuelle Darstellung und Interaktion mit dem IVA sowie der 2D-UI realisiert wird.

Abbildung 4.1 zeigt den realen Aufbau der Versuchsdurchführung sowie eine Fotomontage der in-VR-Situation mit dem intelligenten virtuellen Agenten.



(a) Realer Versuchsaufbau mit VR-Headset und Desktop-PC, auf dem Sprachpipeline und VR-Anwendung laufen (eigene Darstellung)



(b) Fotomontage der sprachbasierten Befragung mit IVA (nach Reallusion Inc., 2025a) und VR-Umgebung (nach Puradiredja et al., 2025) (eigene Darstellung)

Abbildung 4.1: Darstellung des Versuchsaufbaus: links die reale Versuchsanordnung, rechts die VR-Szene mit IVA-Befragung

Die Versuchsdurchführung erfolgt mit einem *Meta Quest 3* Virtual-Reality-Headset, das per PCVR-Verbindung mit einem leistungsstarken Rechner mit moderner Grafikkarte (GPU) verbunden ist. Dadurch wird eine flüssige Echtzeitdarstellung und eine stabile Verarbeitung der Sprachinteraktion gewährleistet. Das zugrunde liegende Sprachmodell sowie die Sprachsynthese werden über externe Dienste bereitgestellt.

Die Datenerhebung erfolgt in einem ruhigen Testraum ohne äußere Ablenkungen oder Umgebungsgeräusche. Die Teilnehmenden stehen während der VR-Erfahrung, um sich frei im virtuellen Raum orientieren zu können. Für die Interaktion innerhalb des VR-Szenarios wird in beiden Bedingungen ein Controller verwendet, der zur Navigation, zur Auswahl von Objekten und zur Bedienung der Benutzeroberfläche dient. In der IVA-Bedingung erfolgt die Beantwortung der Fragen ausschließlich per Spracheingabe, während in der UI-Bedingung der Controller zur Fragebogenerhebung eingesetzt wird. Nach jedem VR-Durchlauf werden die ergänzenden Fragebögen außerhalb der VR an demselben PC ausgefüllt, auf dem auch die VR-Anwendung läuft. Die Versuchssteuerung und Datenerfassung erfolgen zentral über die Sprachpipeline, die sämtliche Interaktionsereignisse, Transkriptionszeiten und Nutzerantworten automatisch protokolliert.

Auf Basis des in diesem Kapitel beschriebenen Forschungsdesigns werden die Systemumsetzung und die Integration der einzelnen Komponenten im folgenden Kapitel detailliert beschrieben.

5 Umsetzung des Prototyps

Zentrales Ziel des Prototyps war es, ein System zu entwickeln, das die sprachbasierte Beantwortung von Fragebögen in einer Virtual-Reality-Umgebung ermöglicht und dabei robuste Kommunikations- und Verarbeitungsprozesse sicherstellt, ohne den Gesprächsfluss durch allzu restriktive Vorgaben einzuschränken. Die Interaktion mit dem Agenten sollte sich möglichst natürlich und flüssig anfühlen.

Die Entwicklung erfolgte in einer modularen Architektur, die die VR-Interaktion, die IVA-Komponente sowie verschiedene Sprachverarbeitungsmodule miteinander verknüpft. Damit bildet der Prototyp die Grundlage für die experimentelle Untersuchung der in dieser Arbeit formulierten Forschungsfragen.

5.1 Systemarchitektur

Die Systemarchitektur des entwickelten Prototypen ist so gestaltet, dass sie eine sprachbasierte Interaktion zwischen Nutzer:in und einem IVA in einer bestehenden Virtual-Reality-Anwendung ermöglicht. Ziel war es, eine Pipeline zu schaffen, die sowohl den klassischen, in-VR-Questionnaire (inVRQ) als auch eine vollständig sprachgesteuerte Befragung mittels IVA mit Echtzeitfeedback unterstützt. Das System besteht aus zwei gekoppelten Teilsystemen, welche in den folgenden Abschnitten detailliert erklärt werden: einem Python-Backend, das die Sprachpipeline und die Zustandsmaschine ausführt, und einem Unity-Frontend, das den IVA und den inVRQ implementiert. Die Kommunikation zwischen dem Frontend und dem Backend erfolgt asynchron über eine WebSocket-Verbindung. In Python wird ein WebSocket-Server gestartet, mit dem sich die Unity-Anwendung verbindet. Die Gesamtarchitektur des Prototyps ist in [Abbildung 5.1](#) dargestellt. Sie zeigt die Verknüpfung der Python-Backend-Komponenten mit dem Unity-Frontend, die Kopplung über WebSockets sowie die Anbindung mit dem HMD.

Das Unity-Frontend wurde für die Studie in eine VR-Anwendung integriert und übernimmt die Darstellung und Animation des IVA. Hier werden Sprachsegmente abgespielt, Texte der Items angezeigt und alle Interaktionen mit dem inVRQ-Fragebogen umgesetzt. Das Python-Backend beinhaltet neben der Sprachpipeline auch Funktionen zum Protokollieren der Metriken und

der Fragebogen-Antworten. Eine zentrale Zustandsmaschine koordiniert die Zustände (*idle*, *listening*, *transcribing*, *thinking*, *synthesizing*, *speaking*; siehe Abb. 5.2) des Systems. In Unity werden diese Zustände für die Steuerung der Animation und Audiowiedergabe gespiegelt.

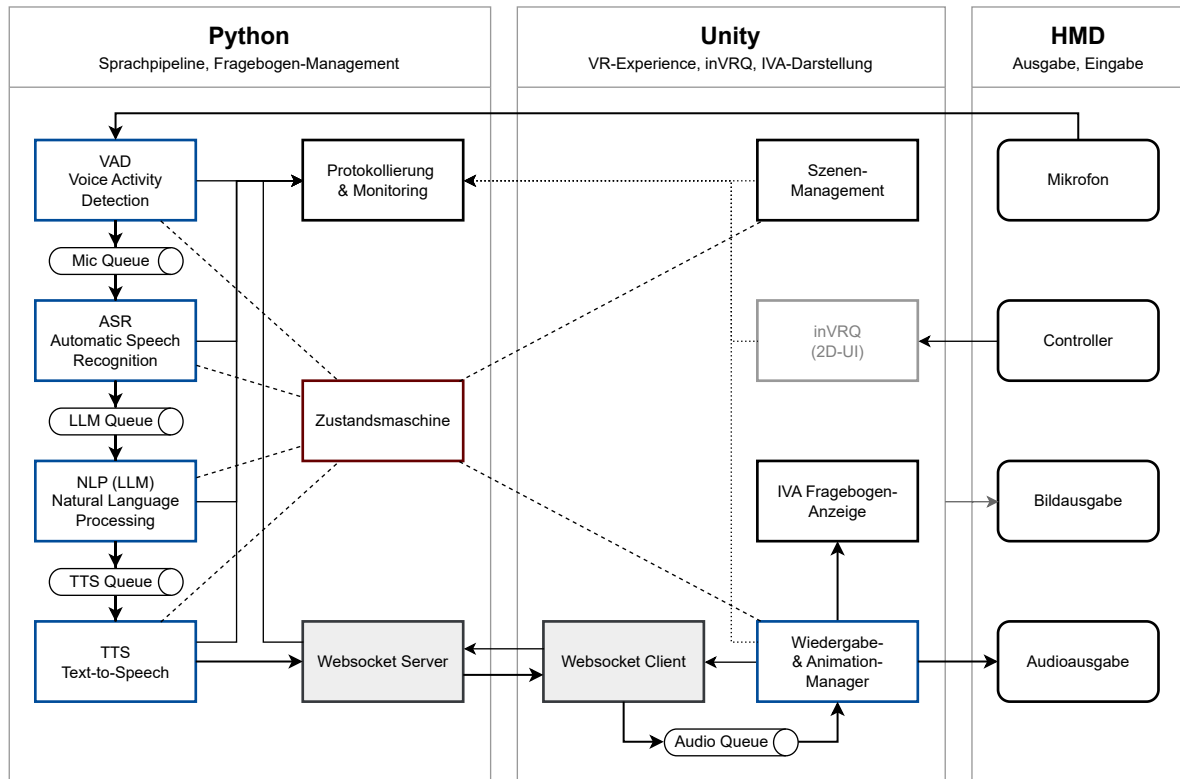


Abbildung 5.1: Übersicht der Systemarchitektur. Die Sprachverarbeitung (VAD, ASR, LLM, TTS) läuft im Python-Backend und kommuniziert über WebSockets mit dem Unity-Frontend, welches die IVA-Darstellung, Fragebogen-UI und Audioausgabe für die Nutzer:innen übernimmt (eigene Darstellung)

5.1.1 Zustandsmaschine

Ein wesentliches Element der Systemarchitektur ist eine Zustandsmaschine (*Finite-State-Machine*, FSM), welche den Ablauf der Sprachpipeline und der IVA-Integration koordiniert. Sie stellt sicher, dass die Verarbeitungsschritte von der Aufnahme des Sprachsignals bis zur Wiedergabe der Agentenantwort klar definierten Regeln folgen und robust ablaufen.

Wie in Abbildung 5.2 dargestellt, beginnt die Maschine im inaktiven Zustand (*Inactive*), bis die Fragebogenerhebung gestartet wird. Danach wird entweder in den Grundzustand (*Idle*) oder mit einem Initialprompt direkt in den *Thinking*-Zustand gewechselt. Aus *Idle* heraus aktiviert die Spracherkennung bei Sprachaktivität den *Listening*-Zustand. Sobald der erste

vollständige Satz vorliegt, wechselt die Maschine in *Transcribing*, wo das Audiosignal in Text umgewandelt wird. Anschließend folgt der *Thinking*-Zustand, in dem das LLM eine Antwort generiert. Über Gabelungen (*Forks*) wird parallel der *Synthesizing*-Zustand gestartet, sobald Teile der Antwort verfügbar sind. Nach dem ersten Audiopakete wechselt die Maschine zusätzlich in den *Speaking*-Zustand, sobald die Wiedergabe beim IVA beginnt. Durch die Fork-Knoten können die Generierung, Synthese und Ausgabe der Antwort überlappen, um die gefühlte Verarbeitungsdauer zu reduzieren. Die Vereinigungsknoten (*Joins*) gewährleisten, dass die Zustände am Ende wieder synchronisiert werden.

Darüber hinaus berücksichtigt die State Machine Unterbrechungen durch die Nutzer:innen: Erfolgt während *Thinking*, *Synthesizing* oder *Speaking* eine neue Eingabe, wird die laufende Verarbeitung über einen separaten Interrupt-Pfad abgebrochen und die Maschine kehrt unmittelbar zu *Listening* zurück.

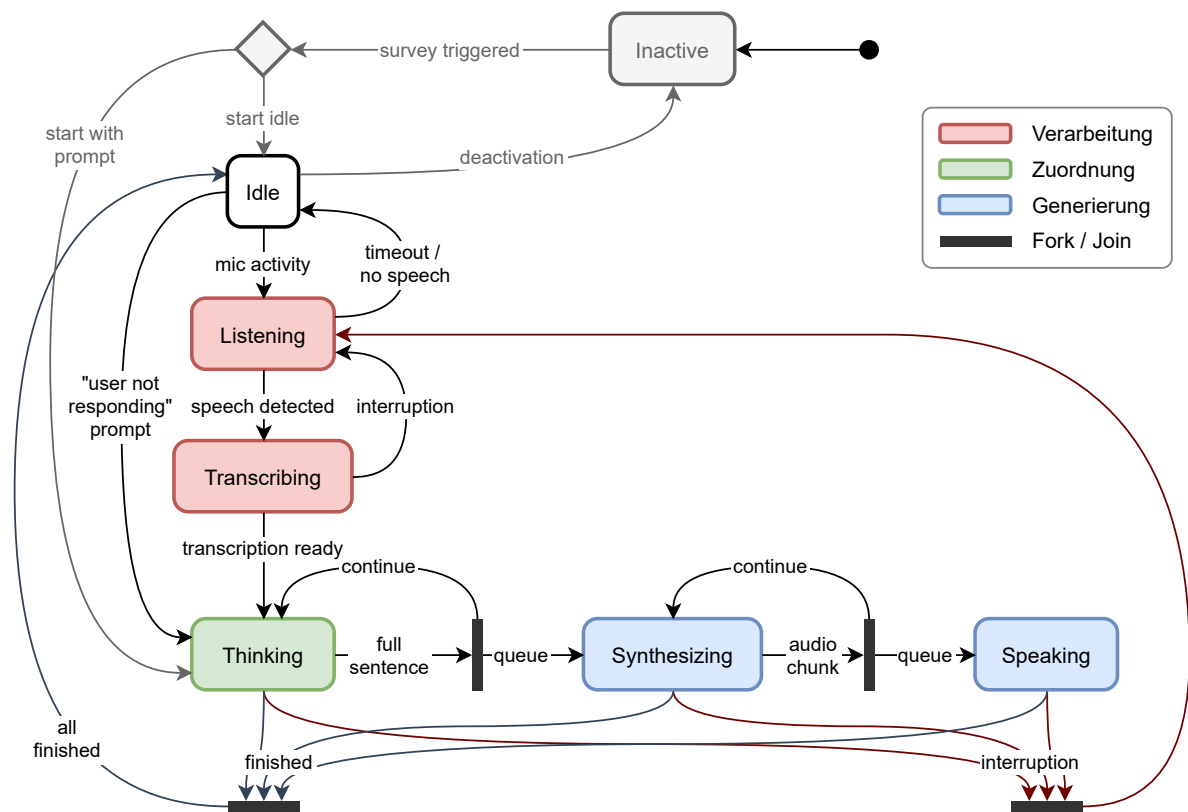


Abbildung 5.2: Zustandsdiagramm der Zustandsmaschine (FSM) zur Steuerung von Spracherkennung, Antwortgenerierung, Sprachsynthese und Wiedergabe (eigene Darstellung)

5.1.2 WebSocket-Kommunikation

Die Kommunikation zwischen dem Unity-Frontend und dem Python-Backend erfolgt asynchron über eine WebSocket-Verbindung. Diese Schnittstelle stellt sicher, dass Statusänderungen, Steuerbefehle sowie Audio- und Textdaten mit geringer Latenz zwischen den beiden Systemen übertragen werden.

WebSocket ist ein Netzwerkprotokoll, das auf TCP (*Transmission Control Protocol*) basiert und eine dauerhafte, bidirektionale Verbindung zwischen Client und Server ermöglicht. Im Gegensatz zu klassischen HTTP-Anfragen, bei denen der Client für jede Anfrage eine neue Verbindung aufbauen und auf eine Antwort warten muss, bleibt die WebSocket-Verbindung nach dem initialen Verbindungsaufbau (*Handshake*) dauerhaft bestehen. Dadurch können beide Seiten – Client und Server – jederzeit Daten senden oder empfangen, ohne dass eine neue Anfrage erforderlich ist. Dieses Prinzip erlaubt eine Echtzeitkommunikation mit geringer Latenz.

Der WebSocket-Server ist Bestandteil des Python-Backends und wird mithilfe des Pakets *websockets* (Augustin & Contributors, 2025) realisiert. Er läuft auf demselben PC wie die übrigen Komponenten der Sprachverarbeitung. Der Server verwaltet eingehende Client-Verbindungen und verteilt Nachrichten an die zuständigen Verarbeitungskomponenten innerhalb der Sprachpipeline.

Nach dem Start der VR-Anwendung baut das Unity-Frontend über die Bibliothek *NativeWebSocket* (Dreyer, 2025) eine persistente Verbindung zu diesem WebSocket-Server auf. Eingehende Nachrichten werden an die zuständigen Komponenten im Unity-Client verteilt, während ausgehende Ereignisse, wie das Abspielen von Sprachaudio-Segmenten, Aktionen (z. B. Beginn oder Ende einer Befragung) und Nutzerantworten aus der 2D-UI, an das Backend übermittelt werden. Bei Verbindungsproblemen zwischen Front- und Backend werden Nachrichten zwischengespeichert und nach Wiederherstellung der Verbindung übertragen.

Die Übertragung selbst ist paketbasiert. Audiodaten aus der TTS-Synthese werden als Binärpakete gestreamt und im Client zu kontinuierlichen Audiosequenzen zusammengesetzt. Ergänzend werden Metadaten wie Zeitstempel und Texte im JSON-Format (JavaScript Object Notation, ein standardisiertes Textformat für strukturierte Daten) gesendet. Auf diese Weise werden auch alle Zustandsänderungen der State Machine an Unity übertragen, sodass Wiedergabe und Animationen synchron gesteuert werden können.

Die WebSocket-Kommunikation bildet das Rückgrat der Integration von VR-Frontend und Sprachpipeline. Sie gewährleistet einen kontinuierlichen Datenfluss, ermöglicht ein enges Zusammenspiel von Audioverarbeitung und Animation und stellt sicher, dass alle Interaktionen zeitnah erfasst und protokolliert werden.

5.2 VR-Szenario

Für das VR-Szenario wurde eine kurze, etwa drei bis fünf Minuten dauernde Anwendung gewählt, die immersiv, aber kognitiv und motorisch nicht fordernd ist. Grundlage bildet ein Modul einer virtuellen Lernumgebung für die Reaktion auf den Ausbruch von Epidemien, die ursprünglich für das Training von Fachkräften im Einsatz mobiler Labore entwickelt wurde (Puradiredja et al., 2025). Für die Zwecke dieser Studie wurde ein Modul gewählt und so vereinfacht, dass keine spezifischen Fachkenntnisse erforderlich sind.

Das VR-Szenario spielt in einem mobilen Laborzelt, wie es typischerweise in entlegenen Regionen ohne geeignete Infrastruktur zum Einsatz kommt (Affara et al., 2021). Es steht in einer offenen Außenumgebung, die durch Palmen, Sandboden, dezente Nebel effekte und verstreut liegende Objekte einen provisorischen Schauplatz vermittelt (siehe Abb. 5.3a). Der Eingangsbereich ist mit Absperrkegeln, Flatterband und einer Plane markiert, wodurch eine klare Abgrenzung zum Umfeld entsteht. In der näheren Umgebung befinden sich zusätzliche Objekte wie Fässer, Kisten oder ein Generator, die den Charakter einer temporären Feldstation verstärken. Im Inneren des Zeltes sind Labortische mit Geräten, Kisten und Probenmaterialien angeordnet, die eine realistische Laboratmosphäre schaffen (siehe Abb. 5.3b). Hinweisschilder wie „Biohazard“ oder „No Access for Unauthorized Personnel“ unterstreichen den thematischen Rahmen der Anwendung.

Die Aufgabe der Teilnehmenden besteht darin, sich in dem mobilen Labor, dargestellt als Zelt, umzusehen und Gegenstände zu identifizieren, die dort nicht vorhanden sein sollten, etwa weil sie eine potenzielle Kontaminationsgefahr darstellen. Die zu findenden Objekte werden für jeden Durchlauf zufällig aus einem vordefinierten Pool ausgewählt, um Abwechslung zu schaffen und Lerneffekte zu minimieren.



(a) Außenansicht des Laborzelts mit abgesperrtem Bereich; Startposition der Teilnehmenden (b) Innenansicht des mobilen Labors mit Arbeitsplätzen und typischer Laborausstattung

Abbildung 5.3: Screenshots des VR-Szenarios in Unity (VR-Umgebung nach Puradiredja et al., 2025; eigene Darstellungen)

Die Teilnehmenden starten im Hauptmenü, das als stilisierte Ansicht des Laborzelts gestaltet ist. Dort erhalten sie eine kurze Einführung in die Steuerung der Anwendung sowie in ihre Aufgabe in diesem Szenario. Anschließend durchsuchen sie das Zelt, visieren verdächtige Objekte mit dem Controller an und wählen diese per Knopfdruck aus. Zu jedem Objekt wird eine Audiosequenz abgespielt, in der erläutert wird, warum das Objekt ungeeignet oder gefährlich ist. In der Originalfassung muss für jedes Objekt zusätzlich eine Multiple-Choice-Aufgabe gelöst werden. Damit die Aufgabe jedoch auch ohne fachliches Vorwissen vollständig bearbeitet werden kann, wurde dieser Schritt für die Studie entfernt. Sobald alle Gegenstände gefunden wurden, erscheint ein Hinweis, dass das Szenario erfolgreich beendet wurde. Die Teilnehmenden werden dann gebeten, sich vor das Zelt zu teleportieren. Dort füllen sie je nach Versuchsbedingung den Fragebogen entweder über die klassische UI oder im Gespräch mit dem virtuellen Agenten aus.

Das Szenario wurde ausgewählt, da der IVA optisch stimmig in die Umgebung integriert werden kann und in dem Szenario bereits sowohl eine 2D-UI als auch gesprochene Texte verwendet werden. Dadurch lassen sich beide Befragungsmodalitäten gleichermaßen in die VR-Umgebung einbetten. Die visuelle Gestaltung der Anwendung ist stilisiert, jedoch realitätsnah, und wird durch binaurale Umgebungsgeräusche ergänzt, die die Immersion verstärken. Die Anwendung ist primär für die Nutzung im Stehen konzipiert, kann jedoch auch sitzend vollständig absolviert werden. Die Interaktion mit der virtuellen Umgebung erfolgt über VR-Controller und ist einfach gehalten. Für die Auswahl von UI-Elementen und Objekten wird der Trigger-Knopf verwendet. Als Fortbewegungsmethode steht ausschließlich Teleportation zur Verfügung. Dafür wird mit dem Joystick eine Art Angel ausgeworfen, mit der festgelegte Ankerpunkte anvisiert werden können (siehe Abb. 5.4). Beim Loslassen des Joysticks teleportiert sich der Avatar zum anvisierten Ankerpunkt.



(a) Interaktion mit einem fehlerhaften Objekt (defekte Kühlbox) innerhalb des Laborzelts (b) Fortbewegung innerhalb der VR-Umgebung durch Teleportation mittels Controller

Abbildung 5.4: Interaktions- und Navigationsmechanismen im VR-Szenario (VR-Umgebung nach Puradiredja et al., 2025; eigene Darstellungen)

5.3 IVA-Integration

5.3.1 Visuelles Design

Für die visuelle Repräsentation des Agenten wurde ein Standard-Charaktermodell aus der Software iClone 8 von Reallusion eingesetzt (Reallusion Inc., 2025a). Das Modell ist in einem realistischen Stil gestaltet und repräsentiert eine männliche Erscheinung (siehe Abb. 5.5). Dementsprechend wurde auch eine männlich klingende Stimme zur Sprachsynthese gewählt. Das Design wurde bewusst neutral und zurückhaltend gewählt, um Voreingenommenheiten zu minimieren und den Fokus der Teilnehmenden nicht von der eigentlichen Interaktion abzulenken.



Abbildung 5.5: 3D-Modell „Kevin“ von Reallusion Inc., 2025a (Quelle: Reallusion Inc., 2025b)

5.3.2 Animationen

Für den Agenten wurden Animationssequenzen für die Zustände Untätig (Idle), Zuhören (Listening und Transcribing), Nachdenken (Thinking und Synthesizing) sowie Sprechen (Speaking) implementiert (siehe Abb. 5.6). Diese werden jeweils von den dazugehörigen Systemzuständen ausgelöst. Pro Zustand steht jeweils nur eine etwa 6 bis 13 Sekunden lange Sequenz zur Verfügung, die bei Bedarf wiederholt abgespielt wird. Für den Idle-Zustand gibt es mehrere längere Animationen, die zufällig ausgewählt werden, um Wiederholungen zu vermeiden. Zusätzlich wird während des Gesprächs Blickkontakt mit den Teilnehmenden gehalten. Im Zustand *Thinking* wird der Blick nach oben links gerichtet, um den Eindruck von kognitiver Verarbeitung zu vermitteln.

5.3.3 Aufbau der Befragung in VR

Nach Abschluss der VR-Erfahrung gelangen die Teilnehmenden durch Teleportation vor den Eingangsbereich des Laborzelts, wo die anschließende Fragebogenerhebung stattfindet. Dieses Setting ist für beide Erhebungsbedingungen identisch und bleibt in derselben Umgebung des zuvor erlebten Szenarios. Abbildung 5.8b zeigt den Aufbau der Befragungsszene vor dem Zelt. Pro Bedingung wird jeweils nur eines der beiden Erhebungssysteme angezeigt – entweder der IVA mit Item-Anzeige oder die 2D-UI.



(a) IVA und Item-Anzeige in der virtuellen Umgebung aus Blickwinkel der Teilnehmenden (b) Position der Teilnehmenden (transparent eingezeichnet) bei der Fragebogenerhebung; pro Bedingung entweder IVA mit Item-Anzeige (mittig) oder 2D-UI (rechts)

Abbildung 5.8: Screenshots der Befragungsszene vor dem Laborzelt in VR (Agent nach Reallusion Inc., 2025a; VR-Umgebung nach Puradiredja et al., 2025; 2D-UI nach Feick et al., 2020; eigene Darstellung)

In der IVA-Bedingung befindet sich der Agent bereits während der gesamten VR-Erfahrung vor dem Zelt, ist aber inaktiv. Sobald sich die Teilnehmenden nach Abschluss des Szenarios an die Position vor dem Zelt teleportieren, wird ein initialer Prompt an das LLM gesendet, wodurch der Agent die Befragung mit einer Begrüßung einleitet.

Neben dem Agenten werden auf einer Anzeigetafel die aktuell gestellte Frage und Antwortanker eingeblendet (siehe Abb. 5.8a). Diese visuelle Unterstützung soll die kognitive Belastung verringern und die Antwortsicherheit erhöhen, indem Teilnehmende die Frage parallel zur auditiven Wiedergabe auch visuell nachvollziehen können. Die Anzeige wird über Metadaten gesteuert, die zusammen mit dem Audiosegment über eine WebSocket-Verbindung vom Python-Backend an Unity übermittelt werden. Diese enthalten den Text des aktuellen Segments sowie Informationen zum jeweiligen Fragebogen-Item. In Unity wird geprüft, ob das Item im aktuellen Segment vorkommt. Wenn es erkannt wird, erscheinen Text und Anker auf der Anzeigetafel. Zu Beginn der nächsten Gesprächsrunde wird die Anzeige wieder geleert.

In der UI-Bedingung erscheint vor dem Zelt eine 2D-UI mit dem klassischen in-VR-Fragebogen (vgl. Abschnitt 5.5). Die Teilnehmenden beantworten die Fragen mit den Controllern über eine *Raycast*-Interaktion, indem sie die jeweiligen Skalenwerte oder Eingabefelder auswählen.

5.4 Sprachpipeline

Die Sprachpipeline bildet das technische Kernstück der Interaktion mit dem IVA. Hier wird die gesprochene Sprache der Nutzer:innen in Echtzeit verarbeitet und in eine hörbare Antwort des Agenten zurückgeführt. Gleichzeitig wird das Ausfüllen des Fragebogens protokolliert und überwacht. Ziel der Pipeline ist es, einen möglichst natürlichen Dialogfluss mit geringen Latenzzeiten und ohne Ausfälle zu ermöglichen sowie einen lückenlosen Datensatz des Fragebogens zu generieren.

Die Pipeline setzt sich aus mehreren aufeinanderfolgenden Modulen zusammen. Zunächst erkennt ein VAD die Sprachaktivität und segmentiert die Mikrofoneingabe. Diese Audiodaten werden anschließend von einer automatischen Spracherkennung (ASR) transkribiert und in Text überführt. Auf dieser Basis generiert ein LLM kontextabhängige Antworten, die schließlich durch ein Text-to-Speech-System (TTS) in Sprache synthetisiert und an Unity zur Wiedergabe übermittelt werden.

Alle Module laufen parallel in separaten Threads und kommunizieren über gemeinsam genutzte Warteschlangen (*Queues*), sodass eine kontinuierliche Verarbeitung möglich ist, ohne dass die einzelnen Module aufeinander warten müssen. Dadurch können etwa Antwortgenerierung, Sprachsynthese und Wiedergabe teilweise parallel erfolgen und die wahrgenommene Latenz reduziert werden. Um lange Ladezeiten beim ersten Nutzereingang zu vermeiden, werden die VAD- und ASR-Modelle zusätzlich beim Start mithilfe eines kurzen Test-Audios vorgeladen.

Im Folgenden werden die einzelnen Komponenten der implementierten Sprachpipeline detailliert beschrieben – von der Audioeingabe und Sprachaktivitätserkennung über Spracherkennung, Sprachmodell und Sprachsynthese bis hin zu Fragebogen-Logik, Turn-Taking und modularer Architektur.

5.4.1 Audioeingang

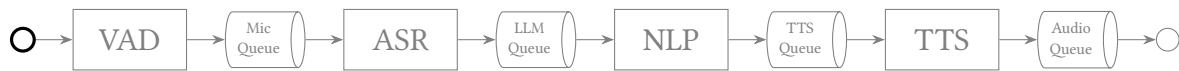


Abbildung 5.9: Module der der Sprachpipeline mit markiertem Audioeingang

Als Eingabegerät wurde in der Studie das in der Meta Quest 3 integrierte Mikrofon genutzt. Es liefert ein vergleichsweise leises, aber ausreichend klares Signal, sodass keine zusätzliche Vorverarbeitung wie Rauschunterdrückung oder Pegelnormalisierung erforderlich war. Die Rohsignale werden direkt in Python mit einer Abtastrate von 16 kHz erfasst und direkt an das VAD-Modul weitergeleitet.

5.4.2 Voice Activity Detection

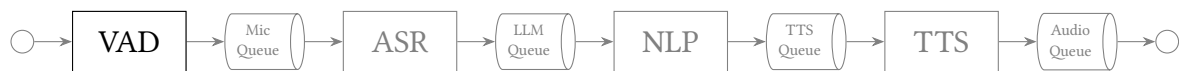


Abbildung 5.10: Module der Sprachpipeline mit markierter Voice Activity Detection

Für die Sprachaktivitätserkennung wurde das Silero-VAD-Modell verwendet, ein leichtgewichtiges vortrainiertes neuronales Netz zur Voice Activity Detection, das als Python-Paket bereitgestellt wird (Silero Team, 2024). Es ist für den Einsatz in Echtzeit optimiert und ermöglicht eine schnelle sowie robuste Erkennung von Sprachsegmenten.

Das Modell wird in Python mithilfe von *PyTorch* (Paszke et al., 2019) geladen, einer Open-Source-Bibliothek zur Entwicklung und Ausführung neuronaler Netze. Der integrierte PyTorch Hub stellt dafür eine Schnittstelle zum Abruf und zur Verwendung vortrainierter Modelle bereit. Anschließend wird das Modell lokal auf der GPU ausgeführt. Für jedes Audiostück wird von dem Modell ein Konfidenzwert errechnet, der gegen einen Schwellwert geprüft wird. In dieser Studie wurde ein Schwellwert von 0,6 (60 % Konfidenz) gewählt, da sich dieser in Tests mit dem verwendeten HMD als geeignete Balance zwischen ausbleibenden und fehlerhaften Auslösungen erwiesen hat. Überschreitet der Konfidenzwert diese Schwelle, wird das Segment als Sprache klassifiziert. Die erkannten Segmente werden zusammen mit einem kurzen Audiopuffer zusammengetragen, um abgeschnittene Satzanfänge zu vermeiden. Sobald der Konfidenzwert für eine festgelegte Zeitspanne (in dieser Studie: 0,5 s) unterhalb des Schwellwerts bleibt, werden die zusammengetragenen Audiosegmente über eine Queue an das ASR-Modul für die automatische Spracherkennung weitergeleitet.

5.4.3 Automatic Speech Recognition

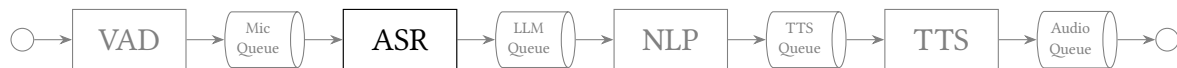


Abbildung 5.11: Module der Sprachpipeline mit markierter Automatic Speech Recognition

Die Umwandlung der erkannten Sprachsegmente in Text erfolgt durch OpenAI Whisper, ein auf der Transformer-Architektur basierendes Encoder-Decoder-Modell (Radford et al., 2022). Für Whisper stehen mehrere multilinguale Modell-Checkpoints unterschiedlicher Größen zur Verfügung, die sowohl eine reine Transkription als auch eine Übersetzung des erkannten Textes in andere Sprachen ermöglichen.

In dieser Arbeit kam das Modell *Whisper-large-v3-turbo* zum Einsatz, eine auf Geschwindigkeit optimierte Variante des *Whisper-large-v3*-Checkpoints. Die Beschleunigung wird durch eine Reduzierung der Decoder-Schichten von 32 auf 4 erreicht. Dadurch verringert sich die Latenz deutlich, während die Genauigkeit nur geringfügig beeinträchtigt wird (Lacombe, 2024). Für die Studie wurde die auf die deutsche Sprache spezialisierte Variante *primeline/whisper-large-v3-turbo-german* verwendet (primeLine AI Services, 2024).

Auch das ASR-Modell wird über den PyTorch-Hub in Python geladen und lokal auf der GPU ausgeführt. Die Verarbeitung erfolgt in einem dedizierten Thread, der kontinuierlich Audio Pakete aus einer Eingabewarteschlange entnimmt. Die Audiosegmente werden zunächst zu einem kontinuierlichen Signal zusammengeführt, bevor sie dem Modell zur Transkription übergeben werden. Treffen während der Verarbeitung weitere Audiosegmente ein, werden diese ebenfalls transkribiert und die Ergebnisse gesammelt. Anschließend wird das vollständige Transkript über eine Warteschlange an die nachgelagerte Sprachverarbeitung weitergeleitet.

Fehlerquellen und Gegenmaßnahmen

Ein zentrales Problem bei der Verwendung von OpenAI Whisper ist das Auftreten von **Halluzinationen** (siehe Abschnitt 2.3.2). Diese treten insbesondere dann auf, wenn das Eingangsaudio keine sprachlichen Anteile enthält, also ausschließlich Stille oder Rauschen vorliegt. In solchen Fällen erzeugt das Modell mitunter vollständige Sätze oder Satzteile, die im Originalaudio nicht vorhanden sind, etwa Phrasen wie „Vielen Dank“, „Untertitelung des ZDF, 2020“ oder „Vertraue und glaube, es hilft, es heilt die göttliche Kraft!“.

Koenecke et al. (2024) zeigen, dass solche Halluzinationen systematisch auftreten können, insbesondere in Situationen mit längeren Pausen oder hohem Rauschanteil. Um dieses Risiko zu minimieren, wird in der vorliegenden Arbeit auf eine möglichst rauschfreie Audioaufnahme gesetzt, was durch das integrierte Mikrofon der Meta Quest 3 gewährleistet ist. Zusätzlich sorgt der eingesetzte VAD dafür, dass sehr kurze Segmente und reine Hintergrundgeräusche

nicht transkribiert werden. Obwohl vereinzelte Halluzinationen nicht vollständig ausgeschlossen werden können, sind diese selten und können daher vernachlässigt werden. Das nachgelagerte LLM ist darüber hinaus instruiert, bei kontextlosen Transkriptionen die Nutzer:innen um eine Wiederholung der eigentlichen Antwort zu bitten.

Ein weiteres Problem betrifft **Fehltranskriptionen** durch Whisper. Besonders häufig tritt dies bei kurzen Audiosignalen auf, in denen nur ein Wort vorkommt. So wird beispielsweise manchmal „Bye“ erkannt, obwohl eigentlich „Zwei“ oder „Drei“ gesagt wurde. Solche Fehler sind insbesondere kritisch, wenn sie vom System als Abbruchsignal interpretiert werden könnten. Daher wurde das LLM explizit instruiert, den Fragebogen nicht vorzeitig zu beenden. Zusätzlich wird im Falle der Transkription „Bye“ eine gesonderte Nachricht an das LLM übermittelt, die auf eine wahrscheinliche Fehltranskription hinweist. Für andere Arten von Transkriptionsfehlern (beispielsweise „Künft“ statt „Fünf“) werden keine besonderen Instruktionen gegeben, da das LLM in den meisten Fällen selbstständig nachfragt oder, vorwiegend bei Freitextantworten, die Eingabe korrekt interpretiert.

5.4.4 Natural Language Processing

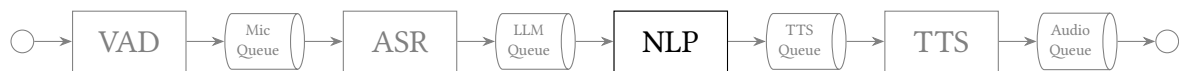


Abbildung 5.12: Module der Sprachpipeline mit markiertem Natural Language Processing

Die semantische Verarbeitung der transkribierten Nutzereingaben erfolgt durch ein LLM. Um eine qualitativ hochwertige und zugleich schnelle Verarbeitung zu ermöglichen, wurde für die Studie eine Cloudlösung eingesetzt. Zum Einsatz kam ein GPT-4.1-Modell von OpenAI (*gpt-4.1-2025-04-14*). Sobald eine neue Transkription in der Eingangswarteschlange erscheint, wird daraus eine Eingabe (*Prompt*) für die Generierung der Agentenantwort erstellt.

Die Kommunikation mit dem Modell erfolgt über die offizielle OpenAI-API, die eine auf HTTPS basierende REST-Schnittstelle bereitstellt. Für die Implementierung wird das Python-Paket *openai* (OpenAI, 2025c) verwendet, das eine asynchrone Anbindung und Echtzeitverarbeitung per Streaming unterstützt. Eingaben werden im JSON-Format übermittelt, und die Antworten werden mit `client.responses.stream` als kontinuierlich gestreamte JSON-Fragmente empfangen.

Das Streaming-Verfahren ermöglicht es, Textsegmente bereits während der Generierung an die Sprachsynthese weiterzuleiten. Dadurch beginnt die Audiowiedergabe oft, bevor die gesamte Antwort vorliegt, was die wahrgenommene Antwortlatenz deutlich reduziert. Bei Verbindungsunterbrechungen oder Fehlern wird der Prompt erneut an das LLM übertragen, sodass der Dialog auch nach kurzzeitigen Netzwerkfehlern fortgesetzt werden kann.

Das LLM übernimmt in diesem Prototypen nicht nur die Generierung der Agentenantworten, sondern stellt auch die Fragen der jeweiligen Fragebogenitems selbst und gibt dabei zusätzlich sowohl die Nummer der gestellten als auch die der beantworteten Frage, sowie die Antwort der Teilnehmenden aus. Dieser Ansatz hat mehrere Vorteile: Zum einen bleiben die möglichen Nutzereingaben uneingeschränkt, ohne dass eine aufwendige regelbasierte Intent-Erkennung erforderlich wäre. So können Teilnehmende etwa nach Erläuterungen der Fragen oder um Korrekturen zuvor gegebener Antworten bitten. Ferner wirkt der Gesprächsfluss natürlicher, da das LLM Bestätigungen und Übergänge selbst formulieren und Rückfragen bei unklaren Antworten stellen kann. Da das LLM jedoch gelegentlich fehlerhafte Angaben ausgibt, werden die Angaben des LLM in einem nachgelagerten Schritt automatisch überprüft und gegebenenfalls korrigiert.

Promptgenerierung und LLM-Konfiguration

Jeder Prompt setzt sich aus mehreren Elementen zusammen. Die Gesamtstruktur ist in Abb. 5.13 dargestellt. Konstant bleiben die Systeminstruktionen sowie das Ausgabeformat. Der Developer-Input wird dynamisch an den aktuellen Kontext angepasst, während der User-Input stets die jeweils aktuelle Transkription enthält.

Die Systeminstruktionen und der Developer-Input sind auf Englisch verfasst. Abhängig von der eingestellten Sprache (für deutsch- oder englischsprachige Teilnehmende) existieren angepasste Varianten der Systeminstruktionen, hinsichtlich der Ausgabesprache und der Beispielantworten.

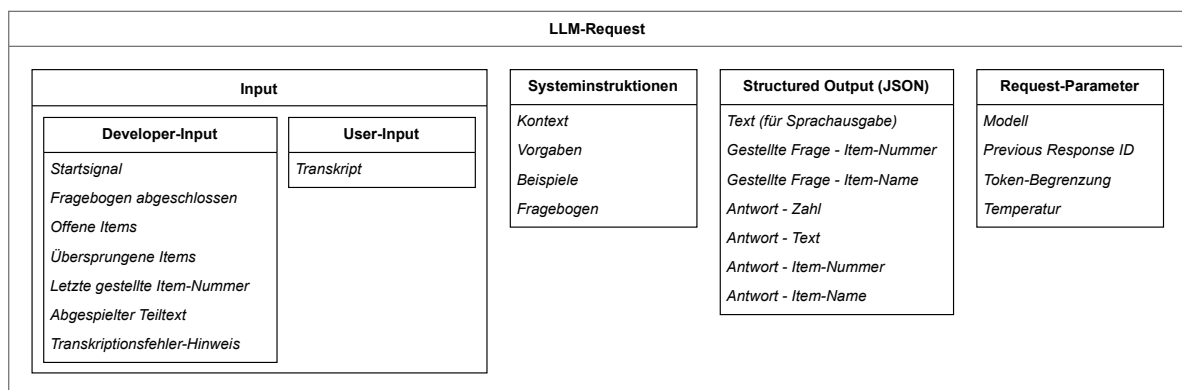


Abbildung 5.13: Struktur des LLM-Requests: links die dynamischen Promptanteile der aktuellen Gesprächsrunde (Developer-Input und User-Input), in der Mitte die statischen Vorgaben des Systems (Instruktionen und vorgegebenes Ausgabeformat), rechts die weiteren Request-Parameter (eigene Darstellung)

Systeminstruktionen: Die Systeminstruktionen enthalten eine Reihe an Anweisungen, die dem LLM den Kontext der Konversation vorgeben, das gewünschte Antwortverhalten spezifizieren und typische Dialogsituationen durch Beispiele illustrieren. Sie definieren die Rolle des Modells als Interviewer, der die Fragebogenitems in einer natürlichen, immersiven Gesprächsweise stellt, Antworten entgegennimmt und den Dialogfluss steuert. Enthalten sind Vorgaben zur Sprache, zum Sprachstil (kurze, TTS-freundliche Sätze, Vermeidung technischer Referenzen), zur Gesprächsstruktur (Einleitung mit Small Talk, systematische Abfolge der Items, kein vorzeitiges Abbrechen) sowie zum Umgang mit Antworten (Validierung und Bestätigung von Likert-Antworten, Akzeptieren textueller Antworten mit eindeutiger Zuordnung zu Likert-Ankern, Nachfragen bei offenen Fragen, Möglichkeit zur Korrektur früherer Antworten). Durch integrierte Beispiel-Dialoge wird das gewünschte Verhalten zusätzlich konkretisiert. Die vollständigen Instruktionen sind in Anhang A.1 dokumentiert.

Ausgabeformat: Das verwendete GPT-Modell unterstützt sogenanntes *Structured Output*, bei dem dem LLM ein exaktes JSON-Ausgabeformat vorgeschrieben wird, an das es sich strikt halten muss. Diese Vorgabe wird durch ein Verfahren namens *Constrained Decoding* umgesetzt, das sicherstellt, dass die Modellgenerierung zu jedem Zeitpunkt einer festgelegten Struktur folgt (Geng et al., 2025). Im Gegensatz zu einem unstrukturierten JSON-Output garantiert dieses Verfahren, dass alle Felder stets syntaktisch korrekt und im vorgegebenen Format ausgefüllt sind. Inhaltliche Fehler können dennoch auftreten und werden daher in der nachgelagerten Datenextraktion überprüft. Das Ausgabeformat ist in Tabelle 5.1 schematisch zusammengefasst; das vollständige JSON-Schema befindet sich in Anhang A.2.

Für das Ausfüllen des Fragebogens umfasst das definierte Schema mehrere Pflichtfelder, darunter die Nummer und Bezeichnung der gestellten und beantworteten Frage sowie die Antwort in numerischer oder freier Textform. Zusätzlich wird ein Feld für den auszugebenden Agententext spezifiziert, der anschließend an die TTS-Komponente wei-

Tabelle 5.1: Ausgabeschema der Modellantworten im Structured Output Format

Feld	Typ	Beschreibung
Text	string	TTS-Ausgabetext
Gestellte Frage – Item-Nummer	integer, null	Nummer des aktuell gestellten Items
Gestellte Frage – Item-Name	string, null	Bezeichnung des aktuell gestellten Items
Antwort – Zahl	integer (1–7), null	Numerische Antwort auf Likert-Item
Antwort – Text	string, null	Freitextantwort
Antwort – Item-Nummer	integer, null	Nummer des beantworteten Items
Antwort – Item-Name	string, null	Bezeichnung des beantworteten Items

tergereicht wird. Alle Felder sind obligatorisch, dürfen jedoch den Wert *null* annehmen, wenn im jeweiligen Kontext keine Angabe vorliegt.

Developer Input: Die Nachricht mit der Rolle *Developer* wird aus mehreren Faktoren, je nach Gesprächsverlauf, zusammengestellt:

- Für die allererste Nachricht wird ein Startsignal gegeben, da das Gespräch vom Agenten initiiert wird.
- Wenn alle Fragen beantwortet sind, enthält der Prompt einen Hinweis, dass der Fragebogen abgeschlossen ist.
- Falls noch Fragen offen sind, werden die fehlenden Items aufgelistet. Übersprungene Fragen werden zusätzlich markiert, damit sie als Nächstes gestellt werden.
- Jede Antwort wird mit einem Verweis auf die zuvor gestellte Frage ergänzt, um den Zusammenhang zwischen der Antwort und der zugehörigen Frage zu kennzeichnen.
- Wenn die letzte Agentenbotschaft durch eine Aussage des Teilnehmenden unterbrochen wurde, wird der bereits abgespielte Teiltext hinzugefügt, damit das LLM weiß, welchen Wissensstand die Nutzer:innen haben.
- Für häufig auftretende Transkriptionsfehler werden spezifische Hinweise ergänzt, dass es sich vermutlich um einen Erkennungsfehler handelt. Beispielsweise wird das Wort „Bye“ als potenzielle Fehltranskription von „Zwei“ oder „Drei“ behandelt.

User Input: Der User Input enthält den transkribierten Text der Nutzer:innen. Falls zuvor mehrere Transkriptionen aufgelaufen sind, die nicht verarbeitet oder abgespielt wurden, werden sie zu einer Nachricht zusammengeführt.

Ein Beispiel für einen Prompt, bei dem zuvor das letzte IPQ-Item (Nummer 13) gestellt wurde, der oder die Nutzer:in mit „Zwei“ geantwortet und dabei die Wiedergabe unterbrochen hat, ist in Codeblock 5.1 dargestellt.

Codeblock 5.1: Beispiel eines User- und Developer-Inputs bei Nutzerunterbrechung

```
1 [{
2   "role":"developer",
3   "content":"Unanswered item numbers: 13, 14, 15, 16
4   The last question you asked was item number 13. The user may be answering this item.
5   Your last text was interrupted by the user, this is the text that was played before
6     ↳ interruption: 'Danke, deine Antwort ist fünf. Jetzt die letzte Frage auf der
7     ↳ Skala: Die virtuelle Welt erschien dir wirklicher als die reale Welt. Von eins'
8   Evaluate if you need to restate the missing part of the last message!"
9 },{
10  "role":"user",
11  "content":"Zwei"
```

Previous Response ID: Zusätzlich wird bei jedem Prompt die Identifikationsnummer (ID) der zuletzt vom LLM erzeugten Antwort übergeben, sodass der bisherige Gesprächsverlauf einbezogen werden kann. Falls frühere Antworten nicht abgespielt wurden, wird nur die ID der zuletzt vollständig oder teilweise wiedergegebenen Nachricht berücksichtigt.

Tokenbegrenzung: Tokens sind die Verarbeitungseinheiten für LLMs; sie entsprechen Textsegmenten, die von einzelnen Zeichen über Teilwörter hin zu vollständigen Wörtern sowie Leer- und Satzzeichen reichen (OpenAI, 2025a). Grundsätzlich ist die Länge der Ausgabe durch das *Structured Output* begrenzt, lediglich Freitextfelder (Nutzerantworten und Agententexte) können in der Länge variieren. In Tests zeigte sich jedoch, dass die GPT-4.1-Modelle gelegentlich mehrere JSON-Blöcke nacheinander erzeugten. Dies führt zu erheblich längeren Verarbeitungszeiten und höheren Kosten. Deshalb wurde eine maximale Tokenanzahl festgelegt, die restriktiv genug ist, um das Erzeugen von übermäßig vielen Ausgaben zu verhindern, aber gleichzeitig ausreichend Spielraum für längere Texte lässt. In dieser Studie wurde ein Wert von 256 Tokens gewählt.

Temperatur: Der Temperaturparameter steuert die Zufälligkeit der Token-Auswahl bei der Generierung. Niedrige Werte führen zu deterministischen und konsistenten Ausgaben, während höhere Werte mehr Varianz erzeugen (Li et al., 2025). Da im Rahmen der Studie primär konsistente und vergleichbare Antworten benötigt wurden, gleichzeitig jedoch repetitive Texte vermieden werden sollten, wurde eine relativ niedrige Temperatur von 0,2 gewählt.

Aus diesem, so zusammengesetzten Prompt generiert das LLM eine passende Antwort im vorgegebenen JSON-Format.

Antwort-Streaming

Die Antwortgenerierung erfolgt im Streaming-Modus: Bereits während das Modell noch Text produziert, werden einzelne Satzfragmente als Teilausgaben direkt über eine Warteschlange an die Sprachsynthese weitergereicht. Dadurch beginnt die Audiowiedergabe oft schon, bevor die gesamte Antwort fertiggestellt ist, was die wahrgenommene Latenz deutlich reduziert. Nach Abschluss des Streamings werden die vom LLM erzeugten Ausgaben aus dem JSON-Objekt extrahiert und ausgewertet. Es erfolgt eine Validierung und ggf. Korrektur der Zuordnung zwischen gestellter Frage und gegebener Antwort und abschließend werden sämtliche Ergebnisse – einschließlich Zeitstempeln, Transkripten, validierten Fragen und Antworten – protokolliert und in Logdateien gespeichert.

Fehlerquellen und Gegenmaßnahmen

Während der Entwicklung des LLM-Moduls für den Prototypen traten verschiedene Probleme auf, die den Gesprächsfluss oder die Korrektheit der Antworten beeinträchtigten. Viele davon ließen sich durch gezieltes Prompt-Engineering abfangen. So enthalten die Instruktionen detaillierte Vorgaben zum Sprachstil, zur Handhabung von Skalenantworten, zum Umgang mit Korrekturen oder Abbruchversuchen sowie zur Wiederholung übersprungener Items. Dennoch blieben einige Fehlerquellen bestehen, die im Folgenden beschrieben werden.

Probleme im Gesprächsfluss

Vorzeitiger Abbruch: Vor allem durch Fehltranskriptionen wie „Zwei“ oder „Drei“ zu „Bye“ kam es dazu, dass das LLM die Befragung fälschlicherweise vorzeitig beendete. Um dies zu verhindern, wurde das LLM in den Systeminstruktionen mehrfach angewiesen, den Fragebogen nicht vorzeitig zu beenden. Für den Spezialfall „Bye“ wurde zusätzlich ein Hinweis mitgegeben, dass es sich vermutlich um eine Fehltranskription handelt.

Nicht wiedergegebene Antworten: Wenn eine Agentenantwort nicht abgespielt wird, könnte das LLM von einem falschen Gesprächsstand ausgehen. Um dies zu vermeiden, wurde bei der *Previous Response ID* ausschließlich die letzte tatsächlich wiedergegebene Antwort berücksichtigt.

Unterbrechungen: Wenn die Nutzer:innen den Agenten unterbrechen, kann es passieren, dass eine Frage nicht vollständig abgespielt wird. Das LLM nimmt dann fälschlicherweise an, dass die Frage vollständig gestellt wurde. Deshalb wird der tatsächlich abgespielte Teiltext im Developer-Prompt ergänzt, sodass das LLM den Gesprächskontext korrekt fortsetzen kann.

Probleme im Modellverhalten

Mehrere JSON-Ausgaben: Teilweise gibt das LLM mehrere JSON-Blöcke in einer Antwort aus. In diesem Fall wird die Antwort als abgeschlossen behandelt, sobald ein vollständiges JSON-Objekt erkannt wird; alle weiteren Tokens werden verworfen.

Übersprungene Fragen: Manchmal überspringt das LLM einzelne Fragen. Deshalb werden im Developer-Prompt alle noch offenen Fragen explizit aufgelistet. Falls Items fehlen, wird das Modell angewiesen, diese gezielt nachzuholen.

Fehlerhafte Identifikationsnummer: In seltenen Fällen wird die ID der vorherigen Unterhaltung nicht von der API (Application Programming Interface) gefunden, was zu Abbrüchen führen kann. Um die Erhebung fortzuführen, wird in diesem Fall der Prompt mit der ID der vorangegangenen Runde erneut an das Modell übergeben.

Probleme in den Modellangaben

Halluzinationen: Wenn dem LLM keine Fragebogen-Items vorliegen, generiert es eigenständig neue Fragen. Dies würde den Ablauf der standardisierten Befragung stören. Um dies zu verhindern, werden in jedem Prompt alle Fragen samt Nummern und Ankern mitgegeben, sodass das Modell ausschließlich aus dieser Liste schöpft. Sobald alle Fragen beantwortet sind, erhält das LLM zudem ein eindeutiges Signal, dass der Fragebogen abgeschlossen ist.

Falsche Fragenzuordnung: Das LLM gibt im Structured Output an, welches Fragebogen-Item im aktuellen Text gestellt wird. Da diese Zuordnung gelegentlich fehlerhaft ist, kommt ein Validierungsverfahren zum Einsatz. Zunächst wird der vom Modell generierte Fragetext mittels unscharfer Suche (Fuzzy-Matching) mit dem im Fragebogen hinterlegten Wortlaut abgeglichen, sodass auch leicht variierte Formulierungen erkannt werden. Wenn keine ausreichende Übereinstimmung gefunden wird, wird in den vorherigen Runden nach einer solchen gesucht. Falls auch so keine ausreichende Übereinstimmung auftritt, wird ein Fehler geloggt, sodass die Zuordnung im Nachgang manuell geprüft werden kann.

Falsche Antwortzuordnung: Da auch die vom Modell zur Nutzerantwort ausgegebene Fragebogen-Item-Nummer manchmal fehlerhaft ist, wird in einem zweiten Schritt überprüft, ob diese mit der in der vorherigen Nachricht gestellten Frage übereinstimmt. Auf diese Weise können Fehlzuordnungen erkannt und korrigiert werden, wodurch eine eindeutige und konsistente Zuordnung aller Antworten zu den jeweiligen Fragen sichergestellt bleibt.

5.4.5 Text-to-Speech



Abbildung 5.14: Module der Sprachpipeline mit markiertem Text-to-Speech-Modul

Der letzte Schritt der Sprachpipeline besteht in der Umwandlung der vom LLM generierten Antworten in gesprochene Sprache. Hierfür wurde das TTS-System von Microsoft eingesetzt (Microsoft, 2025b). Zwar ist die Sprachqualität im Vergleich zu anderen, qualitativ hochwertigeren TTS-Modellen eingeschränkt, diese erfordern jedoch in der Regel eine längere Verarbeitungszeit. Für die technische Implementierung wurde die Open-Source-Bibliothek *rany2/edge-tts* genutzt, welche einen direkten Zugriff auf den Online-TTS-Dienst von Microsoft Edge erlaubt und damit eine einfache Integration in Python-basierte Systeme unterstützt (rany2, 2025).

Die Ansteuerung erfolgt über die asynchrone Schnittstelle der Bibliothek *edge-tts*, die Textdaten zusammen mit Parametern wie Stimme, Sprechgeschwindigkeit und Lautstärke an den Microsoft-TTS-Webdienst übermittelt. Die API erzeugt jeweils den kompletten Audio-Stream einer Eingabe, bevor sie ihn als MP3-Datei zurückgibt. Um die Latenz zu reduzieren, werden die vom LLM erzeugten Antworten satzweise in kleinere Textsegmente aufgeteilt und nacheinander an das TTS-System übergeben. So können die ersten Sprachabschnitte bereits kodiert und abgespielt werden, während das Modell weitere Sätze generiert. Zusätzlich erzeugt Edge-TTS wortgenaue Zeitstempel, sodass nachvollzogen werden kann, wie viel der Agentenantwort bereits wiedergegeben wurde, falls eine Unterbrechung erfolgt.

Für die deutsche Ausgabe wurde die Stimme *de-DE-FlorianMultilingualNeural* eingesetzt. Sie besitzt eine klare Sprachmelodie und gewährleistet eine gute Verständlichkeit im VR-Setting. Für die englische Variante kam *en-GB-RyanNeural* zum Einsatz, da sie eine klassisch britische Akzentfärbung aufweist und in Qualität sowie Sprechgeschwindigkeit der deutschen Stimme ähnelt (Microsoft, 2025a). Die Auswahl fiel bewusst auf neutrale, gut verständliche Stimmen, um Ablenkungen durch starke Emotionen, Dialekte oder unnatürliche Betonungen zu vermeiden.

Die Verarbeitung erfolgt in einem dedizierten Thread, der kontinuierlich Eingaben aus einer gemeinsamen Warteschlange entgegennimmt. Nach der satzweisen Synthese werden die resultierenden Audiosegmente sukzessive über eine WebSocket-Verbindung an Unity übertragen, wo sie in korrekter Reihenfolge abgespielt werden.

Fehlerquellen und Gegenmaßnahmen

Die gewählte deutsche Stimme ist zwar primär für Deutsch ausgelegt, kann aber auch Texte in englischer Sprache ausgeben. In seltenen Fällen trat das Problem auf, dass einzelne deutsche Sätze mit einem englischen Akzent synthetisiert wurden, vermutlich weil die Ausgabe stets nur satzweise erfolgte. Da die verwendete Bibliothek *edge-tts* keine Möglichkeit bietet, die Sprache des Eingabetextes explizit zu fixieren, konnte dieses Verhalten nicht unterbunden werden. Eine spezifische Fehlerbehebung wurde daher nicht implementiert, zumal die Verständlichkeit in allen Fällen ausreichend gegeben war.

5.4.6 Audioausgabe

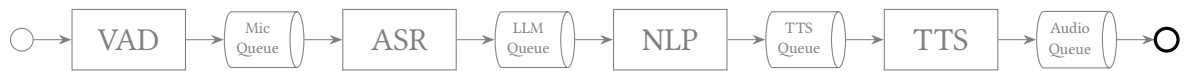


Abbildung 5.15: Module der Sprachpipeline mit markiertem Audioausgang

Die durch das TTS-Modul erzeugten Sprachsegmente werden sequenziell synthetisiert und gestreamt, ohne dass das Ende der Wiedergabe abgewartet werden muss. Die Ausgabe erfolgt strikt nacheinander, sodass sich keine Sprachfragmente überlappen und gleichzeitig möglichst keine Lücken durch die Synthesedauer entstehen. Auf diese Weise ergibt sich eine konsistente und flüssige Audiowiedergabe des IVA.

5.4.7 Fragebogen-Logik und Antwortverarbeitung

Die wichtige Aufgabe der Sprachpipeline besteht darin, den Fragebogen in einem flexiblen und nicht eingeschränkten Gespräch mit den Teilnehmenden auszufüllen. Dazu muss die gesprochene Eingabe nicht nur transkribiert, sondern auch geprüft und so interpretiert werden, dass sie den Fragen des Fragebogens eindeutig zugeordnet und zuverlässig gespeichert werden kann. Damit der Gesprächsfluss nicht durch starre Vorgaben eingeschränkt wird, ist eine robuste Verarbeitung der natürlichen Sprache erforderlich.

Hierfür wird die Transkription an das LLM weitergeleitet, das die Eingabe semantisch auswertet und zusammen mit dem neuen Ausgabertext in einem vordefinierten JSON-Format zurückgibt (siehe Punkt *Ausgabeformat* in 5.4.4). Dieses enthält unter anderem die Nummer der jeweils bearbeiteten Frage sowie den extrahierten Antwortwert in numerischer oder textueller Form.

Dieser LLM-basierte Ansatz erlaubt es den Teilnehmenden, ihre Antworten in eigenen Worten zu formulieren, Zusatzkommentare zu äußern oder Rückfragen zu stellen, ohne dass der Dialogfluss durch feste Antwortmuster begrenzt wäre. Gleichzeitig bleibt die korrekte Auswertung gesichert, da das LLM die relevanten Informationen in strukturierter Form extrahiert und prüft.

Die Steuerung des Fragebogenfortschritts erfolgt weitgehend autonom durch das LLM: Es geht die Fragen in der vorgesehenen Reihenfolge durch und formuliert die Texte in eigenen Worten. Teilnehmende können ihre Angaben jederzeit ändern, indem sie den Agenten um Korrekturen oder Wiederholungen bitten. Auch Rückfragen zu bereits gegebenen Antworten sind möglich. Das System speichert alle Gesprächsrunden mit den gestellten Fragen und gegebenen Antworten ab und überprüft fortlaufend, ob der Fragebogen vollständig bearbeitet wurde. Eingriffe erfolgen nur, wenn Antworten fehlen oder Fragen übersprungen wurden.

Dieser Prozess sorgt für eine dynamische und unbeschränkte Gesprächsführung. So können Teilnehmende beispielsweise ihre Antwort korrigieren, präzisieren („Eigentlich doch eher eine Vier statt eine Fünf.“) oder zusätzliche Erklärungen geben, ohne dass der Erhebungsprozess gestört wird. Damit entsteht eine Balance zwischen der Offenheit eines natürlichen Gesprächs und der strukturierten Datenerfassung, die für die statistische Auswertung erforderlich ist.

Beispielhafter Dialogausschnitt

Zur Veranschaulichung des Ablaufs zeigt der folgende, aus der Pilotstudie entnommene Ausschnitt die Funktionsweise des sprachbasierten Erhebungsprozesses. Der Agent führt durch die Fragen, erkennt die Antworten und reagiert in natürlicher Gesprächsform. Bei Freitextfragen fasst er die Aussage zusammen und stellt gegebenenfalls Nachfragen.

Jede Nachricht des LLMs enthält zudem die Nummer des aktuell gestellten Items sowie die Nummer und die Antwort des zuletzt beantworteten Items. Zu den Nutzernachrichten werden dem LLM noch ergänzende *Developer*-Nachrichten übergeben, die Hinweise enthalten, wie etwa zu noch unbeantworteten oder übersprungenen Items sowie Instruktionen für den Beginn und das Ende der Fragebogenerhebung.

- 1: **Developer:** „Greet the user. Introduce yourself and have some smalltalk. In a few turns ask them if they are ready. Then, if they are ready, ask them the first question.“
- 2: **Agent:** „Hallo und herzlich willkommen! Schön, dass du da bist. Ich werde dir gleich ein paar Fragen stellen. Du kannst ganz normal mit mir sprechen. Wie geht es dir heute?“
asking item: None *answered item:* None *answer:* None
- 3: **User:** „Hallo. Mir geht es gut.“
Developer: „Unanswered items: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16.“
- 4: **Agent:** „Das freut mich sehr! Die Umfrage besteht aus ein paar kurzen Fragen zu deinem Erlebnis. Meistens kannst du mit einer Zahl von eins bis sieben antworten. Am Ende gibt es noch ein paar offene Fragen, bei denen du frei erzählen kannst. Bist du bereit? Sollen wir anfangen?“
asking item: None *answered item:* None
- 5: **User:** „Ja.“
Developer: „Unanswered items: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16.“
- 6: **Agent:** „Super, dann legen wir los! Die erste Frage: In der computererzeugten Welt hattest du den Eindruck, dort gewesen zu sein. Bitte antworte mit einer Zahl von eins – das heißt ‚überhaupt nicht‘ – bis sieben – das heißt ‚sehr stark‘. Wie war dein Eindruck?“
asking item: 0 (G1) *answered item:* None

7: **User:** „Sieben.“

Developer: „Unanswered items: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16. The last question you asked was item number 0. The user may be answering this item.“

8: **Agent:** „Danke, deine Antwort ist sieben. Kommen wir zur nächsten Frage: Ich hatte das Gefühl, dass die virtuelle Umgebung hinter mir weitergeht. Bitte antworte wieder mit einer Zahl von eins – ‚trifft gar nicht zu‘ – bis sieben – ‚trifft völlig zu‘.“

asking item: 1 (SP1) answered item: 0 (G1) answer: 7 (Sieben)

...

36: **Agent:** „Danke, deine Antwort ist zwei. Jetzt kommen wir zu einer offenen Frage, das heißt, du kannst frei antworten. Wie hast du die Bedienbarkeit und Verständlichkeit der VR-Anwendung erlebt? Gab es dabei Momente, in denen du dich frustriert oder gestresst gefühlt hast? Erzähl gern ein bisschen darüber.“

asking item: 14 (Free1) answered item: 13 (REAL4) answer: 2 (Zwei)

37: **User:** „Also die Bedienbarkeit war sehr gut und einfach. Und ... Eigentlich hat mich gar nichts frustriert, es lief alles gut, und gestresst fühlte ich mich auch nicht.“

Developer: „Unanswered items: 14, 15, 16. The last question you asked was item number 14. The user may be answering this item.“

38: **Agent:** „Danke für deine Antwort! Du hast gesagt, die Bedienbarkeit war sehr gut und einfach. Es gab nichts, was dich frustriert oder gestresst hat, und alles lief super. Möchtest du dazu noch etwas ergänzen, oder sollen wir zur nächsten Frage weitergehen?“

asking item: 14 (Free1) answered item: 14 (Free1) answer: „Also die Bedienbarkeit war sehr gut und einfach [...]“

5.4.8 Turn-Taking

In der Pipeline ist ein einfaches Turn-Taking realisiert. Sobald das VAD ausreichend Sprachaktivität erkennt, werden das LLM- und das TTS-System unterbrochen und die laufende Wiedergabe gestoppt. Die Sprecherrolle geht in diesem Moment also auf die Teilnehmenden über. Sobald die Sprachaktivität endet, wird mit möglichst geringer Latenz die Antwort des Agenten generiert und abgespielt, sodass die Sprecherrolle wieder vom Agenten übernommen werden kann.

Zusätzlich ist ein Stille-Timeout implementiert: Bleiben die Teilnehmenden nach einer gestellten Frage länger als fünf Sekunden stumm, wird das LLM erneut mit einem entsprechenden Prompt angestoßen. Auf diese Weise fordert der Agent aktiv zur Beantwortung auf und verhindert, dass das Gespräch ins Stocken gerät.

Nicht umgesetzt wurde die Erkennung sogenannter *Backchannels*. Das System unterscheidet also nicht, ob es sich bei der Sprachaktivität lediglich um kurze Rückmeldungen (z. B. „Mhm“, „Ja“) handelt. Jede erkannte Sprachaktivität führt zu einem Sprecherrollenwechsel.

5.4.9 Modularität und Unabhängigkeit

Die Architektur der Sprachpipeline ist so gestaltet, dass die einzelnen Module unabhängig voneinander ausgetauscht werden können. Für die Verarbeitungsschritte VAD, ASR, LLM und TTS existieren klar definierte Schnittstellen, sodass alternative Modelle oder Dienste mit vergleichbarer Funktionalität ohne größere Anpassungen integriert werden können. Damit ist es etwa möglich, die eingesetzte Spracherkennung oder Sprachsynthese durch leistungsfähigere oder auf spezifische Sprachen angepasste Systeme zu ersetzen.

Die Pipeline ist nicht an die Unity-Umgebung gebunden. Alle Verarbeitungsschritte laufen unabhängig vom VR-Frontend ab. Verbindet sich keine Unity-Anwendung mit dem Websocket-Server, erfolgt die Ein- und Ausgabe direkt über das System-Audio, wobei dieselben Datenstrukturen und Logmechanismen genutzt werden. Diese Unabhängigkeit erleichtert das Testen, Debuggen und Vergleichen einzelner Module und ermöglicht den einfachen Einsatz der Pipeline auch in alternativen Anwendungsfällen.

Kosten der LLM-Nutzung

Für die Kommunikation mit dem GPT-4.1-Modell fallen nutzungsabhängige Kosten an, die von der Menge der verarbeiteten Tokens abhängen. Zum Zeitpunkt der Datenerhebung berechnete OpenAI für das Modell *gpt-4.1-2025-04-14* Preise von 2,00 \$ pro Million Eingabetokens, 0,50 \$ pro Million zwischengespeicherter (*cached*) Eingabetokens sowie 8,00 \$ pro Million Ausgabtokens (Stand: Mai 2025, OpenAI, [2025b](#)).

Ein vollständiger Studiendurchlauf umfasst etwa 15.000 reguläre Eingabetokens, 100.000 zwischengespeicherte Tokens und rund 3000 Ausgabtokens. Daraus ergeben sich Gesamtkosten von etwa 0,104 \$ pro Versuchsdurchgang, was für die geplante Stichprobengröße von 34 Teilnehmenden einem Gesamtbetrag von rund 3,54 \$ entspricht.

5.4.10 Abgrenzung und nicht umgesetzte Funktionen

Die Sprachpipeline konzentriert sich auf die Kernfunktionen für eine zuverlässige und latenzarme Fragebogenerhebung. Einige klassische Aspekte wurden nicht implementiert, da sie für die vorliegende Studie nicht als notwendig erachtet wurden:

Backchannels und prosodische Analyse: Kurze Rückmeldesignale wie „Mhm“ oder „Ja“ werden nicht gesondert erkannt und interpretiert. Ebenso erfolgt keine Auswertung prosodischer Merkmale wie Emotionen, Sprechtempo oder Betonung.

Multimodale Eingaben: Gesten oder weitere nichtsprachliche Kommunikationssignale werden nicht berücksichtigt. Die Interaktion basiert ausschließlich auf Sprache.

Sprachvorverarbeitung: Auf zusätzliche Schritte wie Rauschunterdrückung, automatische Lautstärkeanpassung oder räumliche Filterung wird verzichtet. Die Audiodaten werden unverändert an das VAD-Modul übergeben.

Mehrsprachigkeit: Die Implementierung ist auf Deutsch und Englisch beschränkt. Weitere Sprachmodelle oder Übersetzungsdienste wurden nicht integriert, könnten aufgrund der modularen Architektur jedoch leicht ergänzt werden.

Qualität vs. Latenz: Bei der Auswahl der Modelle wurde bewusst eine Priorisierung auf Echtzeitverarbeitung gelegt, auch wenn dadurch Einbußen in der Erkennungs- oder Sprachsynthesequalität in Kauf genommen werden.

5.5 Fragebogen Benutzeroberfläche (2D-UI)

Für die Vergleichsbedingung – das Ausfüllen des Fragebogens über eine klassische 2D-Benutzeroberfläche (2D-UI) – wurde das Virtual Reality Questionnaire Toolkit (VRQT) von Feick et al. (2020) eingesetzt. Dieses Toolkit ist speziell für die Integration standardisierter Fragebögen in VR-Anwendungen entwickelt worden und unterstützt verschiedene Fragetypen, darunter Likert-Skalen und Multiple-Choice-Formate. Die Elemente werden, wie auch im restlichen VR-Szenario, mit dem VR-Controller per Raycast ausgewählt.

Da das Toolkit jedoch keine Freitextantworten vorsieht, wurde es für die vorliegende Studie entsprechend erweitert. Für die Texteingabe kam das Spatial Keyboard des XR Interaction Toolkits in der Version 3.2.1 zum Einsatz (Unity Technologies, 2025). Zusätzlich wurde das Keyboard um eine Speech-to-Text-Funktion ergänzt: Hält der oder die Nutzer:in während des Sprechens den Mikrofon-Button gedrückt, wird die Eingabe aufgezeichnet und nach dem Loslassen durch ein Whisper-Modell transkribiert (vgl. Kapitel 5.4.3). Das erkannte Transkript wird anschließend automatisch in das Textfeld übernommen. Abbildung 5.16 (rechts) zeigt unten das Spatial Keyboard und darüber das VRQT mit neu integriertem Textfeld.

Da die Fragen der deutschen IPQ-Übersetzung teilweise recht lang formuliert sind, wurde das Layout für die Likert-Skalen so angepasst, dass pro Seite nur zwei Fragen angezeigt werden. Zudem wurden Schriftgröße und Symbolgrößen leicht erhöht sowie UI-Sounds integriert, die auch im restlichen Szenario Verwendung finden. Die Fragen erscheinen in derselben



Abbildung 5.16: Screenshots des Virtual Reality Questionnaire Toolkits (Feick et al., 2020): links die Darstellung zweier Likert-Skala-Items (IPQ), rechts die Eingabe einer offenen Frage über das Spatial Keyboard (Unity Technologies, 2025) (eigene Darstellung)

Reihenfolge wie in der IVA-Variante. Während die IPQ-Items als Pflichtfelder markiert sind, bleiben die Freitextfragen optional.

5.6 Protokollierung und Metrik-Framework

Der Prototyp erhebt umfangreiche Logdaten, die sowohl technische Kennzahlen als auch Nutzerinteraktionen abbilden. Diese Logs dienen einerseits der Evaluation der Nutzerinteraktion (z. B. Qualität und Passung von Fragen und Antworten, Vollständigkeit der Transkripte, Konsistenz der Antworten) sowie der technischen Evaluation (z. B. Messung von Latenzen und Dauer einzelner Verarbeitungsschritte, Erfassung von Fehlerraten und Robustheit der Pipeline). Alle erfassten Daten werden anonymisiert gespeichert, indem Teilnehmende ausschließlich über IDs identifiziert werden.

Das Logging folgt einem Ereignisschema, das die Zustände und Interaktionen erfasst. Dazu gehören etwa der Beginn und das Ende einer Session sowie der Start und der Abschluss der VR-Experience und der Fragebogenerhebung. Für die Sprachinteraktion mit dem Agenten wird für jedes Sprachereignis ein Datenelement angelegt, das während des Durchlaufs der Pipeline sukzessive mit Logdaten gefüllt wird. Erfasst werden unter anderem Sprachereignisse wie der Beginn und das Ende der Sprachaufnahme, Resultate der Spracherkennung, die generierten Prompts und Antworten des LLMs sowie Start und Ende der Sprachausgabe. Zusätzlich werden Nutzerunterbrechungen (Turn-Taking), Fehlermeldungen und gegebenenfalls die Korrekturen der LLM-Angaben dokumentiert.

Die protokollierten Daten werden in tabellarischer Form gespeichert und enthalten neben den Rohwerten auch abgeleitete Kennzahlen wie die Dauer einzelner Fragen oder die gesamte Bearbeitungszeit eines Fragebogens sowie Latenzen der einzelnen Verarbeitungsschritte.

Während der Durchführung steht eine Live-Konsole zur Verfügung, die die laufenden Interaktionen visuell aufbereitet. Hier werden u. a. die Prompts und Antworten des Sprachmodells tabellarisch dargestellt, fehlende Antworten hervorgehoben und Fehlerereignisse unmittelbar kenntlich gemacht.

5.7 Entwicklungsumgebung und Hardware

Das entwickelte System wurde als PCVR-Setup realisiert. Dafür wurde in der Studie ein Meta Quest 3 HMD verwendet, das per Link-Verbindung mit einem Desktop-PC betrieben wurde. Die Meta Quest 3 verfügt über zwei LCD-Panels mit Pancake-Linsen und einer Auflösung von 2064×2208 Pixeln pro Auge bei bis zu 120 Hz Bildwiederholrate. Das Sichtfeld beträgt etwa 110° horizontal. Die Inside-out-Tracking-Sensoren erfassen Position und Bewegung mit sechs Freiheitsgraden (6DoF) und übermitteln diese an Unity. Das integrierte Mikrofon des Headsets wurde für die Sprachaufnahme genutzt.

Diese Konfiguration ermöglicht es, die Rechenlast der Sprachpipeline (insbesondere GPU-beschleunigte Modelle für ASR und VAD) auf dem PC auszuführen, während die Quest 3 primär Rendering, Tracking und Eingabe übernimmt. Das Backend lief auf einem leistungsstarken Desktop-PC mit einer modernen Grafikkarte (GPU). Für die Sprachverarbeitung waren insbesondere Unity und die GPU-beschleunigten Modelle relevant: Silero VAD 6.0 benötigt kaum GPU-Ressourcen (unter 1 GB VRAM), während Whisper *large-v3-turbo-german* ca. 6 GB VRAM belegt. Für das LLM und die TTS-Komponente wurden cloudbasierte Dienste eingesetzt, um eine höhere Modellqualität bei gleichzeitig echtzeitfähiger Performance zu gewährleisten.

Für das LLM wurde das GPT-4.1-Modell von OpenAI eingesetzt, das laut OpenAI gegenüber früheren Echtzeit-Modellen verbesserte Leistungen in den Bereichen Programmieren, Instruktionsbefolgung und Long-Context-Verständnis zeigt. Das Modell kann Eingaben mit bis zu 1 Million Tokens verarbeiten und weist dabei geringere Kosten und Latenzzeiten als seine Vorgängerversionen auf. Neben GPT-4.1 existieren zwei kleinere Varianten, GPT-4.1 mini und GPT-4.1 nano, die dieselben Architekturverbesserungen enthalten, jedoch für geringere Kosten und schnellere Ausführung optimiert sind und eine geringere Genauigkeit aufweisen (OpenAI, 2025b). Für die Studie wurde die GPT-4.1-Variante gewählt, da sie sich während der Entwicklung als bester Kompromiss zwischen Robustheit, Kosten und Latenz erwies.

Die eingesetzten Softwareversionen und die verwendete Hardware sind in Tabelle 5.2 nach Unity-Umgebung, Python-Backend, verwendeten Modellen und physischer Hardware gegliedert.

Tabelle 5.2: Überblick über Entwicklungsumgebung und Hardware

Komponente	Version / Spezifikation
Unity-Umgebung	
Unity Editor	6000.1.12f1 (Technologies, 2025)
XR Interaction Toolkit	3.1.2
OpenXR Plugin	1.15.0
Universal Render Pipeline	17.1.0
uLipSync	3.1.4 (hecomi & Contributors, 2025)
NativeWebSocket	1.1.5 (Dreyer, 2025)
Python Backend	
Python	3.11 (64-bit) (Foundation, 2025)
PyTorch & Torchaudio	2.8.0 (CUDA 12.6 build) (Paszke et al., 2019)
Transformers	4.52.4 (Wolf et al., 2020)
rany2 Edge-TTS	7.2.3 (rany2, 2025)
OpenAI SDK	1.90.0 (OpenAI, 2025c)
WebSockets	15.0.1 (Augustin & Contributors, 2025)
Eingesetzte Modelle	
VAD	Silero VAD 6.0 (Silero Team, 2024)
ASR	Whisper <i>large-v3-turbo-german</i> (primeLine AI Services, 2024)
LLM	GPT-4.1 Release 2025-04-14 (OpenAI, 2025b)
TTS	Edge-TTS (deutsch: <i>de-DE-FlorianMultilingualNeural</i> & englisch: <i>en-GB-RyanNeural</i>) (Microsoft, 2025b)
Hardware	
Head-Mounted Display	Meta Quest 3 (PCVR via Meta Quest Link) (Meta Platforms, Inc., 2023)
Desktop-PC	NVIDIA RTX 3090 (24 GB VRAM), Intel Core i9-11900K, 32 GB RAM, Windows 11
Audio	integriertes Mikrofon und Lautsprecher des HMD

6 Ergebnisse

In diesem Kapitel werden die Ergebnisse der Studie dargestellt. Nach einem Überblick zu Stichprobe und Durchführungszeiten folgen die Befunde zum Präsenzerleben (IPQ), zu den in VR erhobenen Freitextantworten, zur subjektiven Arbeitsbelastung (Raw NASA-TLX) und zur Gebrauchstauglichkeit der Erhebungsformen (SUS). Abschließend werden die Ergebnisse des Abschlussfragebogens sowie technische Leistungskennzahlen und Fehler des Systems berichtet und Zusammenhänge der verwendeten Messgrößen betrachtet.

Zitate aus den Freitextangaben wurden zur besseren Lesbarkeit orthografisch und grammatikalisch bereinigt, inhaltlich jedoch unverändert übernommen.

6.1 Überblick und Stichprobe

An der Studie nahmen insgesamt 36 Probanden teil. Das Durchschnittsalter lag bei 37,7 Jahren ($SD = 12,8$) mit einer Altersspanne von 20 bis 70 Jahren. In der Stichprobe waren 10 Teilnehmende weiblich, 25 männlich und eine Person nicht binär. Zwei Teilnehmende führten die Befragung auf Englisch durch, während die übrigen 34 auf Deutsch sprachen. Beide Versuchsbedingungen – (A) die sprachbasierte Befragung mit dem IVA und (B) die klassische Fragebogenerhebung über eine 2D-UI – wurden gleichmäßig ausbalanciert: Jeweils 18 Teilnehmende starteten mit Bedingung A oder Bedingung B.

Die Teilnehmenden unterschieden sich stark in ihrer Vorerfahrung mit Virtual Reality (siehe Tab. 6.1). Drei Personen (8,3 %) hatten keinerlei VR-Erfahrung, während zehn Personen (27,8 %)

Tabelle 6.1: Erfahrung der Teilnehmenden mit Virtual Reality

Kategorie	<i>n</i>	%	
Keine Erfahrung	3	8,3	■
Weniger als 5× genutzt	10	27,8	■■■■■
Gelegentlich genutzt	8	22,2	■■■■■
Häufig genutzt	4	11,1	■■■
Besitze eigenes VR-Headset	11	30,6	■■■■■■■

Tabelle 6.2: Erfahrung der Teilnehmenden mit Sprachassistenten

Kategorie	Klassische Sprachassistenten		KI-basierte Sprachsysteme	
	n	%	n	%
Noch nie	7	19,4	14	38,9
Einmal ausprobiert	8	22,2	10	27,8
Gelegentlich genutzt	13	36,1	8	22,2
Regelmäßig genutzt	4	11,1	3	8,3
Täglich im Einsatz	4	11,1	1	2,8

weniger als fünfmal ein VR-System genutzt hatten. Acht Teilnehmende (22,2 %) berichteten von einer gelegentlichen Nutzung, vier (11,1 %) gaben an, VR häufig zu verwenden, und elf Teilnehmende (30,6 %) besaßen ein eigenes VR-Headset.

Auch im Umgang mit klassischen Sprachassistenten wie Alexa, Siri oder Google Assistant zeigten sich deutliche Unterschiede (siehe Tab. 6.2, links): Während 19,4 % noch nie einen Sprachassistenten verwendet hatten, gaben 36,1 % an, diese gelegentlich zu nutzen. Jeweils vier (11,1 %) gaben an, regelmäßig oder sogar täglich Sprachassistenten zu verwenden.

Die Erfahrungen mit KI-basierten Sprachsystemen wie ChatGPT mit Spracheingabe (siehe Tab. 6.2, rechts) waren insgesamt geringer ausgeprägt: Rund 38,9 % hatten bisher keine Erfahrung, weitere 27,8 % hatten ein solches Sprachsystem lediglich einmal ausprobiert. Nur eine Person (2,8 %) berichtete von täglicher Nutzung.

Die Selbsteinschätzung der Persönlichkeit verteilte sich weitgehend gleichmäßig zwischen introvertierten, extravertierten und situationsabhängigen Typen und zeigte keine signifikanten Zusammenhänge mit den übrigen Variablen.

Die Durchführungszeiten unterschieden sich zwischen den Bedingungen deutlich (siehe Tab. 6.3 und Abb. 6.1). Während die VR-Erfahrung selbst in beiden Bedingungen ähnlich lang ausfiel, zeigte sich bei der Fragebogenerhebung ein signifikanter Unterschied: Die Beantwortung der Fragen mit dem IVA nahm durchschnittlich erheblich mehr Zeit in Anspruch (8 : 52 min) als mit der UI (5 : 02 min). Ein Wilcoxon Signed-Rank Test bestätigte diesen Unterschied als signifikant ($z = 8,5$, $p < ,001$, $n = 36$). Insgesamt verbrachten die Teilnehmenden im Mittel rund 43 Minuten mit der Studie, wobei etwa die Hälfte dieser Zeit auf das Beantworten der Fragebögen außerhalb der VR entfiel.

Alle Datensätze lagen vollständig vor; es war kein Ausschluss aufgrund von Abbrüchen oder fehlenden Angaben erforderlich. In drei Fällen musste die Sprachpipeline neu gestartet werden: aufgrund eines nicht erkannten Mikrofons (P01, P13) vor Beginn der Erhebung sowie wegen eines temporären Softwarefehlers während der Datenerfassung (P02).

Tabelle 6.3: Dauer der VR-Erfahrungen und Fragebogenerhebungen

Phase	Bedingung	<i>M</i>	Min	Max	<i>SD</i>
VR-Experience	IVA	3:39	2:03	6:53	1:24
	UI	3:50	1:49	9:03	1:39
Fragebogenerhebung in VR	IVA	8:52	6:24	11:28	1:16
	UI	5:02	2:29	9:45	2:12
Außerhalb der VR		22:09	10:54	44:49	7:40
Gesamt		43:32	23:39	81:58	14:11

Anmerkung: Alle Zeitangaben im Format M:SS (Minuten:Sekunden).

Die betroffenen Sitzungen konnten nach kurzer Unterbrechung fortgesetzt werden, ohne dass die Teilnehmenden das HMD abnehmen mussten. Es kam zu keinen Datenverlusten.

Insgesamt ergeben sich damit für die IPQ- und SUS-Fragebögen jeweils 72 Datensätze. Der Raw-NASA-TLX-Fragebogen wurde getrennt für Skalen- und Freitextfragen erhoben, wodurch sich 2×72 Datensätze ergeben. Darüber hinaus liegen 36 Abschlussfragebögen mit Likert-Items sowie mehrere Freitextangaben vor, die für die qualitative Analyse herangezogen wurden. Ergänzend wurden technische Logdaten (Fehlerraten, Latenzen, Zeitmessungen) für beide Bedingungen aufgezeichnet.

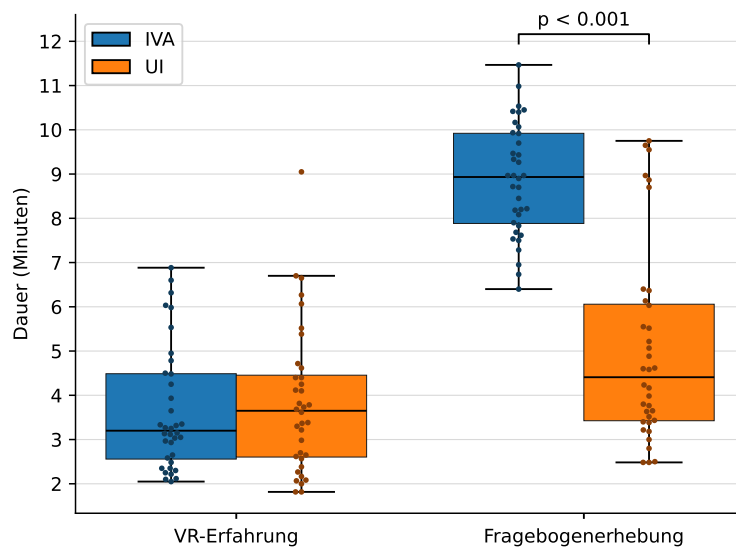


Abbildung 6.1: Boxplots und Einzelwerte der Durchführungszeiten für VR-Erfahrung und Fragebogenerhebung pro Bedingung (IVA, UI). (Verteilungen in Minuten)

6.2 Präsenzerleben (IPQ)

Das Präsenzempfinden wurde mit dem *Igroup Presence Questionnaire* (IPQ) erfasst, der *Allgemeines Präsenzempfinden* (G1), *Räumliche Präsenz* (SP), *Involviertheit* (INV) und *Erlebten Realismus* (REAL) abbildet. Der Fragebogen wurde direkt im Anschluss an die VR-Experience innerhalb von VR erhoben – je nach Bedingung per Sprachinteraktion mit dem IVA oder über eine klassische 2D-UI. Die Antworten erfolgten auf einer siebenstufigen Likert-Skala (1 = niedrige, 7 = hohe Ausprägung) und wurden zur Auswertung auf 0 bis 6 transformiert.

Abbildung 6.2 zeigt die Verteilungen der IPQ-Scores in beiden Bedingungen. Zu erkennen ist, dass sich die Form und Streuung der Verteilungen über die Subskalen hinweg weitgehend ähneln und nur geringe Verschiebungen auftreten.

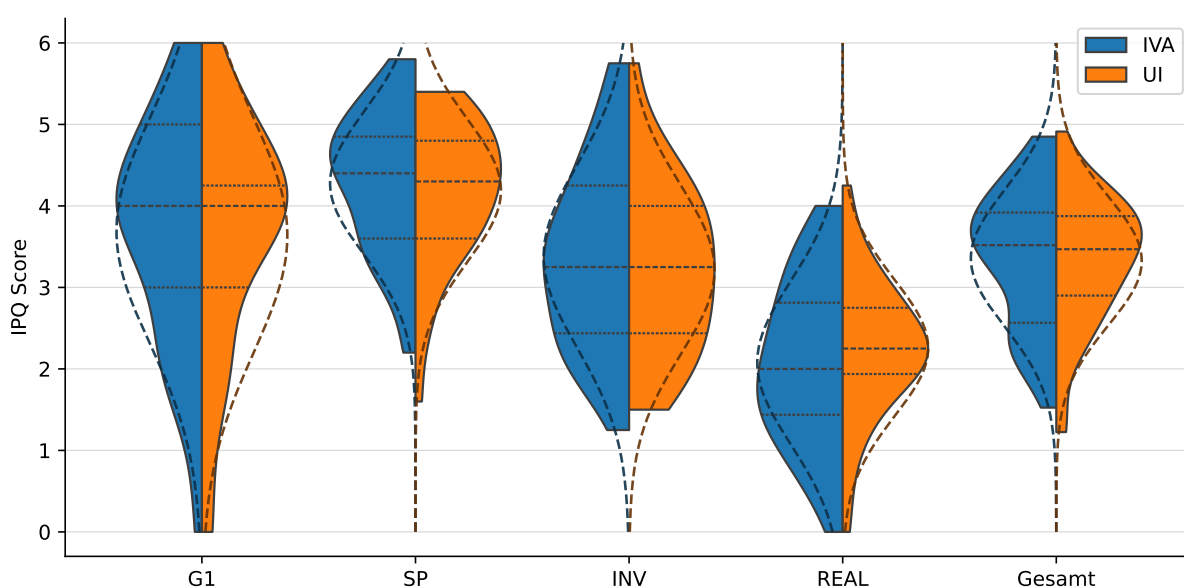


Abbildung 6.2: Violinplots der Verteilungen der IPQ-Subskalen- und Gesamtscores (14 Items) in den beiden Erhebungsbedingungen (eingezeichnet sind Quartile sowie theoretische Normalverteilungen (gestrichelte Linien))

Streuung und Varianzvergleich

Zur Beurteilung der Streuung der IPQ-Werte zwischen den beiden Erhebungsformen wurden Standardabweichungen (SD) sowie das Varianzverhältnis berechnet (vgl. Abschnitt 4.5). Werte von $VR > 1$ deuten auf größere Streuung in der IVA-Bedingung hin. Zur Prüfung signifikanter Varianzunterschiede wurde der Pitman-Morgan-Test für gepaarte Daten durchgeführt. Mithilfe dieses Tests wurde geprüft, ob sich die Varianzen zweier abhängiger Stichproben systematisch unterscheiden. Die Ergebnisse sind in Tabelle 6.4 dargestellt.

Tabelle 6.4: Dispersion und gepaarter Varianzvergleich (IPQ)

Subskala	SD _{IVA}	SD _{UI}	VR	<i>t</i>	<i>p</i>
G1	1,411	1,457	0,94	-0,221	,827
SP	0,885	0,891	0,99	-0,053	,958
INV	1,143	1,049	1,19	0,710	,482
REAL	1,019	0,850	1,44	1,550	,130
Gesamt (14 Items)	0,848	0,804	1,11	0,508	,615

Anmerkung: Freiheitsgrade $df = 34$; VR = Varianzratio s_{IVA}^2/s_{UI}^2

Deskriptiv zeigten sich höhere Streuungen in der IVA-Bedingung für die Subskalen *INV*, *REAL* und den Gesamtscore (*Gesamt*), während *G1* und *SP* nahezu gleiche Standardabweichungen aufwiesen. Die höchsten relativen Unterschiede traten bei der Subskala *REAL* auf (VR = 1,44). Der Pitman-Morgan-Test ergab jedoch für keine Subskala signifikante Varianzunterschiede zwischen den Erhebungsformen ($p \geq ,130$).

Verteilungsform (Schiefe)

Zur Beurteilung der Verteilungsform wurde die Schiefe je Subskala in beiden Erhebungsbedingungen berechnet (vgl. Abschnitt 4.5). Zusätzlich wurde die Schiefe-Differenz $\Delta\gamma = \gamma_{IVA} - \gamma_{UI}$ mit einem 95 %-Konfidenzintervall (Bootstrap, 10.000 Stichprobenwiederholungen) ausgewiesen (siehe Tab. 6.5).

Hinsichtlich der Verteilungsform zeigten sich überwiegend nur geringe Abweichungen von der Symmetrie. Die Subskalen *G1* und *SP* wiesen in der UI-Bedingung jedoch signifikante Abweichungen von der Normalverteilung auf (*G1*: $\gamma_{UI} = -0,91$, $p = ,025$; *SP*: $\gamma_{UI} = -0,85$, $p = ,034$). Für alle Subskalen umfassten die Konfidenzintervalle den Wert 0, was auf keine signifikanten Unterschiede in der Schiefe zwischen den Bedingungen hinweist.

Tabelle 6.5: Schiefe γ und Schiefe-Differenzen der IPQ-Scores zwischen Erhebungsformen

Subskala	γ_{IVA}	$p_{\gamma,IVA}$	γ_{UI}	$p_{\gamma,UI}$	$\Delta\gamma$	95 % Konfidenzintervall
G1	-0,52	,176	-0,91	,025	0,389	[-0,54; 1,33]
SP	-0,48	,208	-0,85	,034	0,373	[-0,64; 1,18]
INV	0,32	,392	0,27	,479	0,057	[-0,69; 0,86]
REAL	0,05	,892	-0,28	,457	0,329	[-0,78; 1,34]
Gesamt (14 Items)	-0,36	,345	-0,73	,066	0,370	[-0,62; 1,20]

Vergleich der IPQ-Scores zwischen den Erhebungsformen

Zur Überprüfung der IPQ-Scores zwischen den beiden Erhebungsformen wurden Mittelwerte, Standardabweichungen sowie inferenzstatistische Vergleichsmaße berechnet (siehe Abb. 6.3 und Tab. 6.6). Ziel war es, zu prüfen, ob sich das subjektive Präsenzerleben zwischen den beiden Erhebungsarten unterscheidet.

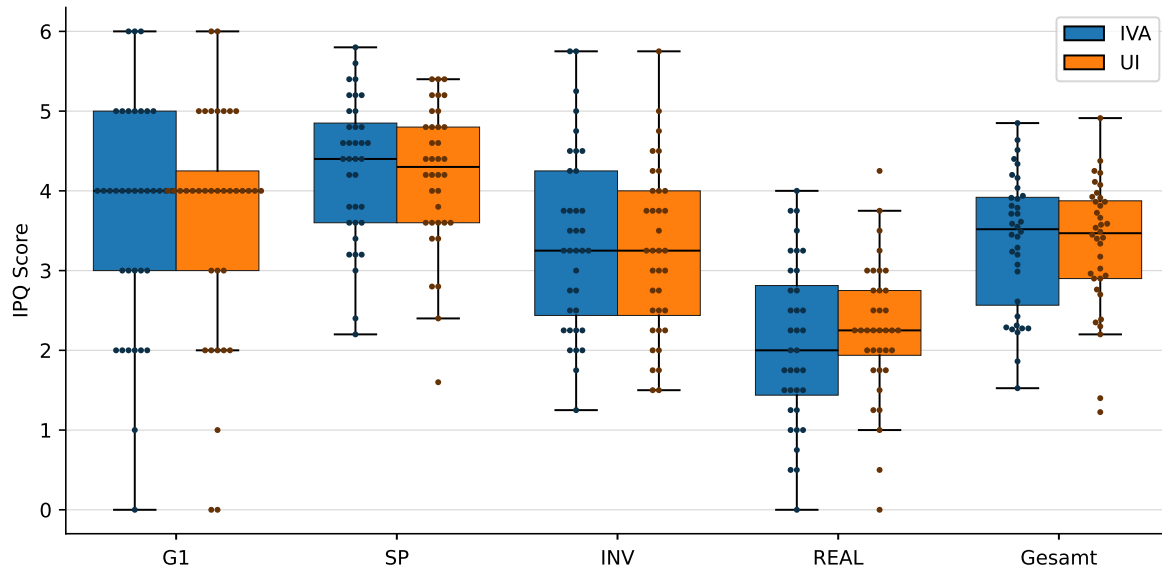


Abbildung 6.3: Boxplots der IPQ-Subskalen- und Gesamtscores (14 Items) der IVA- und UI-Bedingungen (Dargestellung von Mittel- und Einzelwerten)

Die statistischen Vergleiche erfolgten mithilfe gepaarter t -Tests (vgl. Abschnitt 4.5). In keiner der Subskalen zeigte sich ein signifikanter Unterschied zwischen den beiden Erhebungsformen. Ergänzend berechnete Effektstärken (Cohen's d) lagen für alle Dimensionen im niedrigen Bereich ($|d| < 0,3$), was auf geringe bis vernachlässigbare Mittelwertunterschiede hinweist. Die Ergebnisse zeigen, dass sich die beiden Erhebungsbedingungen in keiner der IPQ-Subskalen signifikant unterschieden. Die Mittelwerte lagen in beiden Bedingungen leicht über der Skalenmitte, was auf insgesamt moderate bis leicht positive Präsenzbewertungen hinweist.

Tabelle 6.6: Mittelwerte und Standardabweichungen für das subjektive Präsenzerleben (IPQ)

Subskala	IVA		UI		$t(35)$	p	Cohens d
	M	SD	M	SD			
G1	3,69	1,41	3,64	1,46	0,24	,815	0,04
SP	4,28	0,88	4,19	0,89	0,74	,466	0,12
INV	3,38	1,14	3,22	1,05	1,20	,239	0,20
REAL	2,08	1,02	2,26	0,85	-1,53	,134	-0,26
Gesamt (14 Items)	3,36	0,85	3,32	0,80	0,37	,715	0,06

6.3 Freitextfragen zur VR-Erfahrung

Ergänzend zum Likert-Fragebogen wurden im Anschluss an die VR-Erfahrungen offene Fragen in VR gestellt. Die Teilnehmenden konnten dabei ihre Eindrücke zur *Bedienbarkeit und Verständlichkeit der VR-Anwendung*, zu möglichen *körperlichen Beschwerden (Motion Sickness)* sowie allgemeine *sonstige Anmerkungen zum VR-Erlebnis* äußern. Ziel war es dabei weniger, die Inhalte im Detail zu analysieren, sondern vielmehr zu prüfen, ob sich die Art und Ausführlichkeit der Freitexteingaben in Abhängigkeit von der Bedingung unterschied.

Die Freitextantworten wurden inhaltlich vorverarbeitet, in Einzelaussagen zerlegt und thematisch kodiert. Auf dieser Basis wurden sowohl die *Wortanzahl* als auch die *Anzahl kodierter Themen* quantitativ verglichen. Die Kodierung erfolgte halbautomatisch mittels LLM-basierter Segmentierung und anschließender Plausibilitätskontrolle.

Veranschaulichung der Inhalte

Zur inhaltlichen Einordnung zeigen Abbildung 6.4 zwei Wordclouds mit den häufigsten kodierten Themen pro Bedingung, zusammengefasst über alle drei Freitextfragen. Insgesamt wurden 119 unterschiedliche Themen kodiert, davon wurden 96 in der IVA-Bedingung und 71 in der UI-Bedingung genannt.



(a) Sprachinteraktion mit dem IVA



(b) Texteingabe in der UI

Abbildung 6.4: Wordclouds der kodierten Themen der Freitextantworten, zusammengeführt für alle Fragen

Die Wordclouds und Tabelle 6.7 zeigen, dass in der IVA-Bedingung geringfügig häufiger positive Aspekte der Interaktion wie Verständlichkeit, Bedienbarkeit oder Spaß genannt wurden. In der UI-Bedingung wurden dagegen eher Herausforderungen im Zusammenhang mit der Steuerung (z. B. Controller-Bedienung, Intuitivität) sowie teilweise mit physischen Eigenschaften des VR-Headsets (z. B. „Headset drückt“) angesprochen, was jedoch insgesamt nur als leichte Tendenz zu interpretieren ist.

Tabelle 6.7: Häufigste kodierte Themen (Gesamt und pro Bedingung)

Thema	n	IVA	UI	Thema	n	IVA	UI
Keine Motion Sickness	63	32	31	Interaktion wenig	6	4	2
Kein Stress oder Frust	45	23	22	Interessant	5	4	1
Keine Beschwerden	30	22	8	Verständlich	5	2	3
Gute Bedienbarkeit	18	11	7	Zweite Durchführung	4	2	2
Eingewöhnung	10	4	6	Intuitiv	4	1	3
Gute Verständlichkeit	8	6	2	Immersion gering	4	2	2
Spaß	8	5	3	Keine probleme	4	3	1
Einfach	8	5	3	Navigation unklar	3	3	0
Katze	7	5	2	Positiv	3	2	1
Agent	7	7	0	Immersion	3	2	1
Headset drückt	6	3	3	HMD visuelles Artefakt	3	3	0

6.3.1 Wortanzahl der Antworten

Zur Untersuchung möglicher Unterschiede in der Länge der Freitextantworten zwischen den Bedingungen wurde die durchschnittliche Wortanzahl pro Teilnehmendem verglichen (vgl. Abb. 6.5). Die Analyse erfolgte mithilfe des Wilcoxon-Signed-Rank-Tests (vgl. Abschnitt 4.5). Zur Einschätzung der Effektstärken wurde der Korrelationskoeffizient r berechnet.

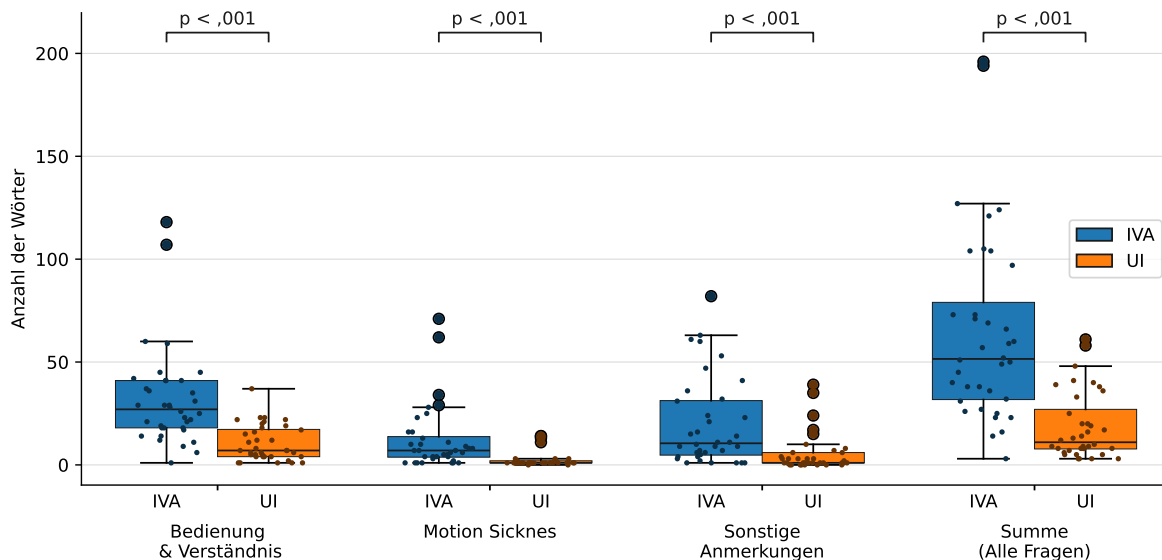


Abbildung 6.5: Wortanzahl der Freitextantworten in den offenen Fragen des inVRQ je Bedingung (Signifikanzniveaus basieren auf Wilcoxon-Signed-Rank-Tests)

Abbildung 6.5 und Tabelle 6.8 zeigen, dass die Teilnehmenden in der IVA-Bedingung in allen drei Fragen sowie in der Gesamtsumme signifikant längere Antworten gaben ($p < ,001$). Die Effektstärken wiesen r -Werte zwischen ,73 und ,87 auf und konnten somit als groß eingestuft werden.

Tabelle 6.8: Wortanzahl in den Freitextantworten pro Antwort

Frage	IVA		UI		n	p	r
	M	SD	M	SD			
Bedienbarkeit / Verständlichkeit	31,89	24,17	10,47	8,68	34	<,001	-,87
Motion Sickness	12,36	15,94	2,53	3,84	30	<,001	-,76
Offene Rückmeldungen	20,19	21,62	5,53	9,41	33	<,001	-,73
Summe (alle Fragen)	64,44	45,83	18,53	16,35	35	<,001	-,87

Anmerkung: n = Paare mit nicht-null Differenz

6.3.2 Anzahl kodierter Themen

Analog wurde die Anzahl der kodierten Themen pro Antwort untersucht. Insgesamt wurden 119 unterschiedliche Themen kodiert, davon wurden 96 in der IVA-Bedingung und 71 in der UI-Bedingung genannt. Abbildung 6.6 und Tabelle 6.9 zeigen, dass in der IVA-Bedingung durchschnittlich mehr Themen genannt wurden. Signifikante Unterschiede bestanden bei *Bedienbarkeit / Verständlichkeit* ($p < ,001$, $r = -,85$) und den *offenen Rückmeldungen* ($p = ,002$, $r = -,68$). Bei *Motion Sickness* ergab sich kein signifikanter Unterschied ($p = ,166$).

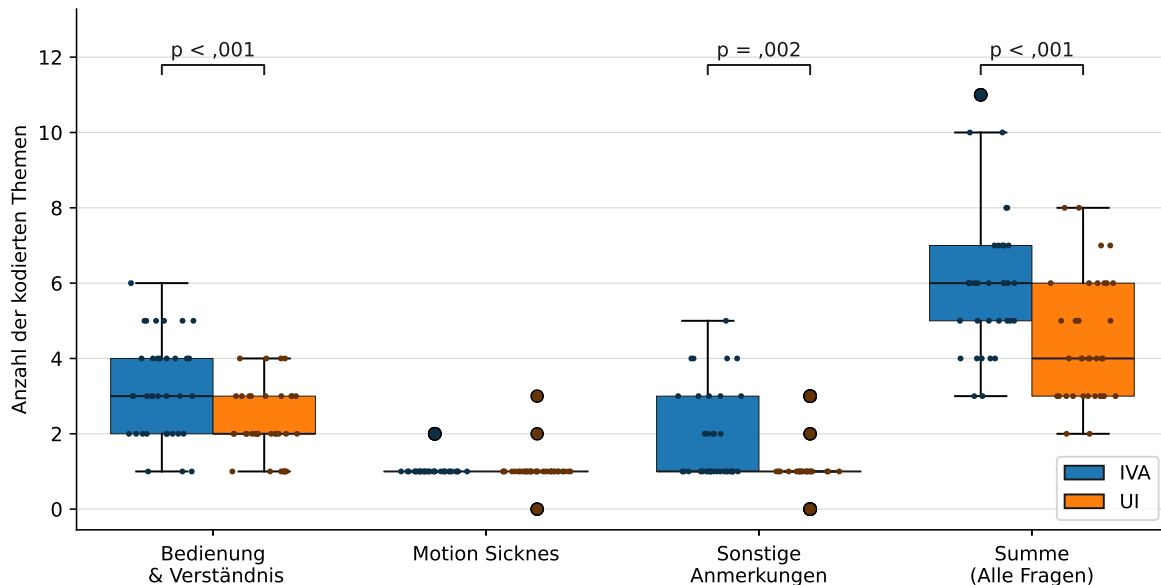


Abbildung 6.6: Anzahl kodierter Themen in den offenen Fragen (IVA vs. UI). (Signifikanzniveaus basieren auf Wilcoxon-Signed-Rank-Tests)

Auch die sprachliche Ausführlichkeit fiel in der IVA-Bedingung höher aus. Exemplarisch: „Aber die Bedienbarkeit war sehr intuitiv, sehr einfach. Keine Probleme.“ (P9144, IVA) vs. „Sehr intuitiv und einfach. Keine Probleme.“ (P9144, UI). Im Mittel verwendeten die Teilnehmenden in der IVA-Bedingung mehr als doppelt so viele Wörter pro kodiertem Thema. Sie nutzten im Durchschnitt etwa 10 Wörter pro Thema ($M_{IVA} = 10,21$, $SD_{IVA} = 8,20$), Teilnehmende in der UI-Bedingung dagegen durchschnittlich nur rund 4 Wörter ($M_{UI} = 3,73$, $SD_{UI} = 2,44$).

Tabelle 6.9: Anzahl kodierter Themen in den Freitextantworten pro Antwort

Frage	IVA		UI		<i>n</i>	<i>p</i>	<i>r</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Bedienbarkeit / Verständlichkeit	3,25	1,30	2,25	0,94	22	<,001	-,85
Motion Sickness	1,17	0,38	1,03	0,45	10	,166	-,44
Offene Rückmeldungen	1,94	1,24	1,22	0,99	21	,002	-,68
Summe (alle Fragen)	6,36	2,15	4,50	1,66	32	<,001	-,77

Anmerkung: *n* = Anzahl der Paare mit nicht-null Differenzen

6.3.3 Antworthäufigkeit und inhaltliche Qualität

Ergänzend wurde der Anteil inhaltlich gehaltvoller Antworten pro Bedingung erfasst. Als *inhaltlich* galten Äußerungen, die über ein einfaches Verneinen oder eine leere Antwort hinausgingen.

Tabelle 6.10 zeigt, dass Teilnehmende in der IVA-Bedingung häufiger inhaltliche Rückmeldungen gaben (60,2 %) als in der UI-Bedingung (47,2 %). Zudem beantworteten in der UI-Bedingung weniger Personen die Fragen zur *Motion Sickness* und *sonstigen Anmerkungen* überhaupt.

Tabelle 6.10: Antworthäufigkeiten und Anteil inhaltlicher Antworten pro Bedingung und Frage

Frage	Bedingung	Antworten	Inhaltliche Antw.	Anteil (%)
Bedienbarkeit & Verständlichkeit	IVA	36	33	91,7
	UI	36	29	80,6
Motion Sickness	IVA	36	9	25,0
	UI	34	5	13,9
Sonstige Anmerkungen	IVA	36	23	63,9
	UI	28	17	47,2
Summe	IVA	108	65	60,2
	UI	98	51	47,2

Anmerkung: *n* = 36; „Inhaltliche Antw.“ bezeichnen Äußerungen, die über ein einfaches Verneinen oder eine leere Antwort hinausgehen.

6.4 Subjektive Arbeitsbelastung (NASA-TLX)

Die subjektiv wahrgenommene Arbeitsbelastung wurde mit dem *Raw NASA-TLX* pro Bedingung zweimal erhoben: einmal bezogen auf die Beantwortung von Likert-Skalen und einmal auf Freitextfragen. Dies ermöglicht eine differenzierte Betrachtung der empfundenen Belastung sowohl in Abhängigkeit von der Befragungsbedingung als auch vom Fragetyp. Analysiert wurden jeweils die Unterschiede zwischen den beiden Bedingungen, nicht jedoch innerhalb einer Bedingung zwischen Skalen- und Freitextfragen, da entsprechende Vergleiche nicht Teil der Forschungsfragen waren und zusätzliche Tests die Teststärke der Hauptanalysen, trotz Holm-Bonferroni-Korrektur, reduziert hätten.

Da die Normalverteilungsannahme in den meisten Dimensionen verletzt war, wurden zur Auswertung Wilcoxon-Vorzeichen-Rang-Tests herangezogen. Die Signifikanzniveaus wurden gemäß der Holm-Bonferroni-Methode korrigiert (vgl. Abschnitt 4.5).

Abb. 6.7 zeigt die Verteilungen und Unterschiede der Arbeitsbelastung in beiden Bedingungen (IVA, UI), getrennt nach Fragetyp. Eingezeichnet sind die signifikanten Unterschiede zwischen den Bedingungen. Tab. 6.11 fasst die deskriptiven Kennwerte und Testergebnisse zusammen.

Insgesamt zeigte sich in der UI-Bedingung eine höhere wahrgenommene Belastung beim Beantworten von Freitextfragen, insbesondere in den Dimensionen *Körperliche Anforderung*, *Leistung* und *Frustration*. Bei den Likert-Skalen hingegen berichteten Teilnehmende in der IVA-Bedingung eine höhere Belastung, vor allem in der Dimension *Frustration*.

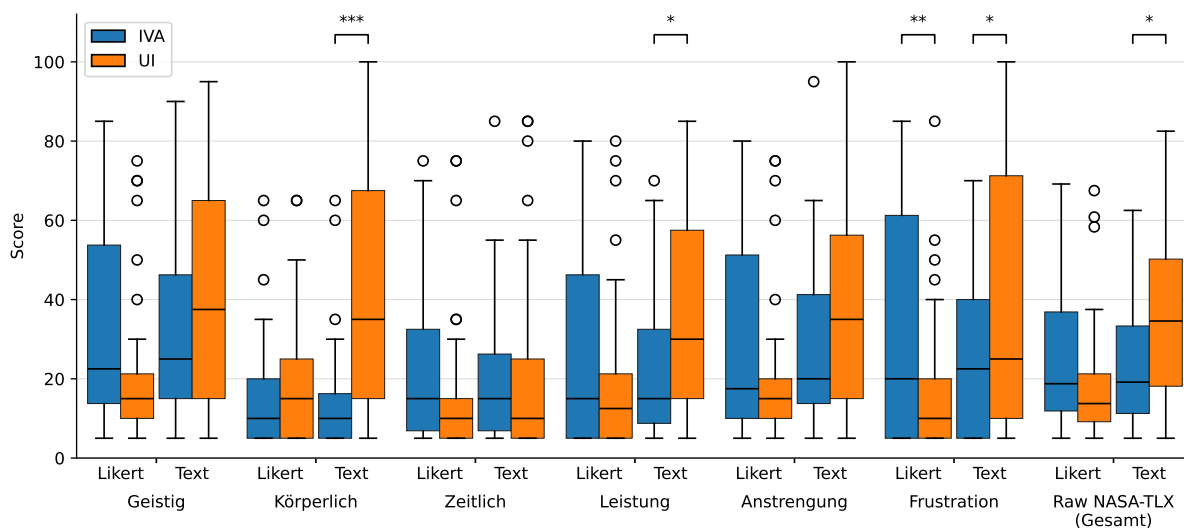


Abbildung 6.7: Verteilung der NASA-TLX-Werte für Likert-Skalen und Freitextfragen in beiden Bedingungen (Boxplots mit Ausreißern), Signifikanzniveaus beziehen sich auf Holm-Bonferroni-korrigierte Tests (* $p < ,05$; ** $p < ,01$; *** $p < ,001$)

Tabelle 6.11: Ergebnisse des Raw NASA-TLX pro Bedingung für Likert-Skalen und Freitextfragen

Dimension	IVA		UI		<i>W</i>	<i>z</i>	<i>p</i>	<i>p</i> _{Holm}	<i>r</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>					
Likert-Skalen									
Geistige Anforderung	32,5	24,8	22,8	19,4	137,5	-2,17	,030	n.s.	-,39
Körperliche Anforderung	14,9	15,3	19,0	16,4	92,5	-1,65	,098	n.s.	-,34
Zeitliche Anforderung	23,7	21,5	16,2	19,0	106,5	-1,99	,047	n.s.	-,38
Leistung	25,0	22,7	20,0	20,3	133,5	-0,79	,432	n.s.	-,16
Anstrengung	30,1	26,0	20,4	19,6	104,5	-1,82	,069	n.s.	-,36
Frustration	32,1	28,1	16,2	17,9	33,5	-3,48	<,001	,007	-,70
Gesamt (Raw)	26,4	19,8	19,1	15,6	128,0	-2,54	,011	n.s.	-,45
Freitextfragen									
Geistige Anforderung	31,7	22,0	44,0	26,6	194,0	-1,98	,047	n.s.	-,34
Körperliche Anforderung	14,1	14,7	42,6	31,2	18,5	-4,21	<,001	<,001	-,79
Zeitliche Anforderung	19,7	17,6	20,7	23,8	159,5	-0,08	,935	n.s.	-,02
Leistung	23,3	18,9	36,9	26,1	99,0	-3,10	,002	,022	-,55
Anstrengung	27,8	21,6	41,0	29,6	132,5	-2,27	,023	n.s.	-,41
Frustration	26,2	21,4	40,1	32,4	77,0	-2,87	,004	,041	-,54
Gesamt (Raw)	23,8	15,6	37,6	23,8	103,5	-3,16	,002	,019	-,55

Anmerkung: *p*_{Holm} = Signifikanzwert nach Holm-Bonferroni-Korrektur; n.s. = nicht signifikant

Für die Likert-Skalen zeigte sich ein signifikanter Unterschied in der Dimension *Frustration* ($W = 33,5$, $p < ,001$, $p_{\text{Holm}} = ,007$, $r = -,70$), wobei die Werte in der IVA-Bedingung deutlich höher ausfielen. Alle weiteren Dimensionen unterschieden sich nach Holm-Korrektur nicht signifikant.

Bei den Freitextfragen zeigten sich insgesamt höhere Werte in der UI-Bedingung. Der Gesamtwert des Raw NASA-TLX war hier signifikant höher als in der IVA-Bedingung ($W = 103,5$, $p = ,002$, $p_{\text{Holm}} = ,019$, $r = -,55$).

Auch auf Ebene einzelner Dimensionen ergaben sich signifikante Unterschiede: Teilnehmende bewerteten insbesondere die *Körperliche Anforderung* in der UI-Bedingung deutlich höher ($W = 18,5$, $p < ,001$, $p_{\text{Holm}} < ,001$, $r = -,79$). Darüber hinaus lagen die Werte in den Dimensionen *Leistung* ($W = 99,0$, $p = ,002$, $p_{\text{Holm}} = ,022$, $r = -,55$) und *Frustration* ($W = 77,0$, $p = ,004$, $p_{\text{Holm}} = ,041$, $r = -,54$) signifikant höher in der UI-Bedingung. Alle übrigen Dimensionen unterschieden sich nicht signifikant zwischen den Bedingungen.

6.5 System Usability (SUS)

Die Gebrauchstauglichkeit der beiden Befragungsbedingungen wurde mit der *System Usability Scale* (SUS) erfasst. Der SUS umfasst zehn Items, die zu einem Gesamtscore zwischen 0 und 100 verrechnet werden (Brooke, 1996). Zur Interpretation der Ergebnisse wird auf die Normdaten von Bangor et al. (2008) Bezug genommen. Sie schlagen vor, Produkte mit Werten unter 70 als „bestenfalls mittelmäßig“ einzustufen. Werte im Bereich der hohen 70er bis 80er deuten auf eine gute bis sehr gute Usability hin, während Scores über 90 als exzellent gelten können.

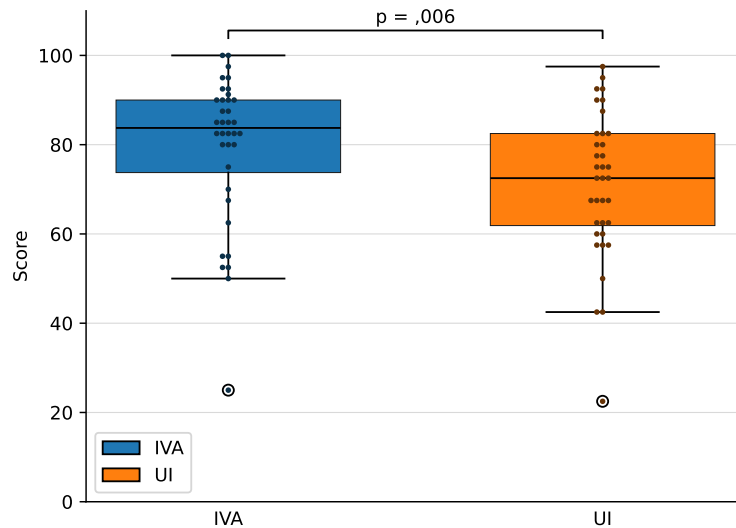


Abbildung 6.8: System Usability Scale (SUS)-Score für IVA und UI (Boxplots mit Einzelwerten)

Da die Normalverteilungsannahme nicht in beiden Bedingungen erfüllt war, erfolgte der Vergleich der beiden Erhebungsformen mittels Wilcoxon-Signed-Rank-Test (vgl. Abschnitt 4.5). Wie Tab. 6.12 und Abb. 6.8 zeigen, lagen die Werte in beiden Bedingungen über dem Referenzwert von 70 und damit im akzeptablen bis guten Bereich. Die IVA-Bedingung wurde jedoch signifikant besser bewertet als die UI-Bedingung ($W = 117,0$, $n = 32$, $p = ,006$, $r = -0,49$), was einem großen Effekt entspricht.

Tabelle 6.12: System Usability Scale (SUS) in beiden Bedingungen

Messwert	IVA		UI		W	n_0	z	p	r
	M	SD	M	SD					
SUS-Score	79,34	16,82	71,04	16,36	117,0	32	-2,75	,006	-,49

6.6 Reliabilität der Erhebungsinstrumente

Zur Beurteilung der internen Konsistenz der eingesetzten Fragebögen wurde Cronbachs α berechnet (vgl. Abschnitt 4.5). Dieses Maß beschreibt die Homogenität der Items einer Skala und gibt Aufschluss über die Zuverlässigkeit der jeweiligen Erhebungsinstrumente.

Tab. 6.13 fasst die Gesamtwerte der drei Hauptinstrumente zusammen; eine detaillierte Aufschlüsselung der Subskalen findet sich im Anhang (Tab. C.1).

Der **IPQ** erzielte in der IVA-Bedingung eine insgesamt gute interne Konsistenz ($\alpha = ,813$) und in der UI-Bedingung eine akzeptable Konsistenz ($\alpha = ,770$). Innerhalb der Subskalen zeigten *Involvement* ($\alpha_{IVA} = ,79$; $\alpha_{UI} = ,60$) und *Realism* ($\alpha_{IVA} = ,82$; $\alpha_{UI} = ,71$) zufriedenstellende bis gute Werte, während die Subskala *Spatial Presence* mit $\alpha_{IVA} = ,54$ und $\alpha_{UI} = ,56$ in beiden Bedingungen eine geringe Homogenität aufwies. Die Itemanalyse zeigte, dass insbesondere das Item *SP3* („Ich hatte nicht das Gefühl, in dem virtuellen Raum zu sein.“; *hatte nicht das Gefühl – hatte das Gefühl*) die interne Konsistenz der Subskala *Spatial Presence* leicht reduzierte ($\Delta\alpha \approx ,01$ bis $,10$).

Der **NASA-TLX** zeigte eine sehr hohe interne Konsistenz über alle Varianten hinweg. Für die Likert-Skalen ergaben sich $\alpha_{IVA} = ,92$ und $\alpha_{UI} = ,91$, für die Freitextfragen $\alpha_{IVA} = ,89$ und $\alpha_{UI} = ,91$. Auch die über beide Fragetypen aggregierten Werte zeigten eine exzellente Reliabilität ($\alpha_{IVA} = ,91$, $\alpha_{UI} = ,92$). Damit weist der NASA-TLX in allen Bedingungen eine sehr homogene Erfassung der subjektiven Arbeitsbelastung auf.

Die **SUS** zeigte in der IVA-Bedingung eine gute interne Konsistenz ($\alpha = ,845$) und in der UI-Bedingung eine akzeptable Konsistenz ($\alpha = ,763$). Damit erweist sich die Skala als zuverlässig zur Erfassung der wahrgenommenen Gebrauchstauglichkeit in beiden Bedingungen.

Tabelle 6.13: Interne Konsistenz (Cronbach's α) der eingesetzten Fragebögen in beiden Bedingungen (Gesamtwerte)

Fragebogen	Bedingung	Cronbach's α	Interpretation
IPQ (Gesamt)	IVA	,813	Gut
	UI	,770	Akzeptabel
NASA-TLX (Gesamt)	IVA	,907	Exzellent
	UI	,924	Exzellent
SUS (Gesamt)	IVA	,845	Gut
	UI	,763	Akzeptabel

Insgesamt weisen die eingesetzten Instrumente durchweg akzeptable bis sehr gute interne Konsistenzen auf. Der IPQ zeigte insgesamt gute Reliabilitäten, wenngleich die Subskala *Spatial Presence* in beiden Erhebungsformen nur eine geringe interne Konsistenz aufwies. Der NASA-TLX erreichte in allen Varianten exzellente Reliabilitätswerte und erweist sich damit als äußerst konsistent. Auch die SUS zeigte eine gute (IVA) bzw. akzeptable (UI) interne Konsistenz. Damit kann die Messqualität der eingesetzten Fragebögen insgesamt als hoch eingeschätzt werden.

6.7 Ergebnisse des Abschlussfragebogens

Im Anschluss an die VR-Erfahrungen beantworteten die Teilnehmenden einen Abschlussfragebogen. Dieser umfasste Einschätzungen zum allgemeinen Wohlbefinden, zur empfundenen Vertraulichkeit der Antworten, zur Verständlichkeit der Fragen sowie zur Vertrauenswürdigkeit des IVA. Einige Items wurden getrennt für Skalen- und Freitextfragen gestellt, um mögliche Unterschiede zwischen den Fragetypen erfassen zu können. Ziel war es, die subjektive Wahrnehmung des Befragungsszenarios und die Präferenzen der Teilnehmenden möglichst umfassend zu erfassen.

Abb. 6.9 zeigt die Antwortverteilungen der Likert-Items in den Themenbereichen Wohlbefinden, Vertraulichkeit, Verständlichkeit, mentale Anforderung und Vertrauenswürdigkeit. Wie Tab. 6.14 zeigt, lagen die Mittelwerte der Antworten überwiegend im neutral bis positiven Bereich.

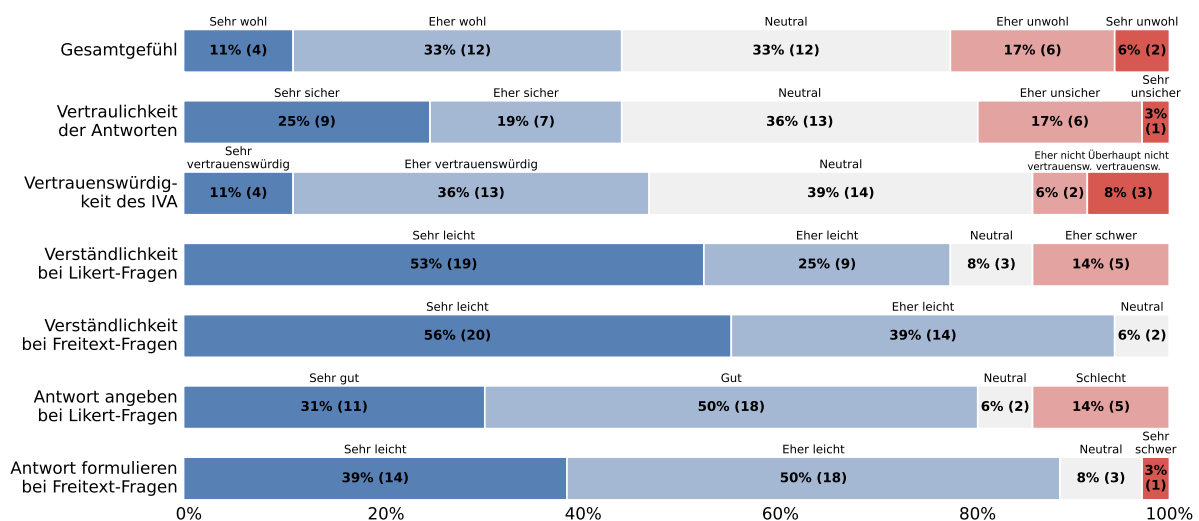


Abbildung 6.9: Antwortverteilungen im Abschlussfragebogen ($n = 36$) (prozentuale Häufigkeiten der gewählten Antwortkategorien; absolute Häufigkeiten in Klammern)

Tabelle 6.14: Mittelwerte der Antworten im Abschlussfragebogen

Frage	Antwortskala	<i>M</i>	<i>Median</i>	<i>SD</i>
Gesamtgefühl	Unwohl – Wohl	0,28	0	1,06
Vertraulichkeit der Antworten	Unsicher – Sicher	0,47	0	1,13
Vertrauenswürdigkeit des IVA	Nicht vertr. – Vertrauenswürdig	0,36	0	1,05
Verständlichkeit Likert-Fragen	Schwer – Leicht	1,17	2	1,08
Verständlichkeit Freitext	Schwer – Leicht	1,50	2	0,61
Antwort angeben (Likert)	Schlecht – Gut	0,97	1	0,97
Antwort formulieren (Freitext)	Schwer – Leicht	1,22	1	0,83

Anmerkung. $n = 36$; Skala von -2 (bzw. „sehr negativ“) bis +2 („sehr positiv“)

Im Folgenden werden die einzelnen Themenbereiche sowie die dazugehörigen Freitextantworten näher beschrieben.

6.7.1 Gesamtgefühl

Auf die Frage nach dem allgemeinen Wohlbefinden gaben insgesamt 33 % der Teilnehmenden an, sich wohl zu fühlen, und 11 % sogar „sehr wohl“, während 33 % neutral antworteten. Rund ein Viertel der Teilnehmenden berichtete dagegen, sich eher (17 %) oder sehr (6 %) unwohl gefühlt zu haben.

In den Freitextantworten wurden vor allem zwei Aspekte wiederholt thematisiert: die **Unge-
wohntheit** der Interaktion mit einer KI sowie die **Unnatürlichkeit bzw. Unangenehmheit** des Agenten. Mehrere Teilnehmende beschrieben, dass es sich zunächst befremdlich anfühlte, mit einer KI zu sprechen:

„Fühlt sich komisch an, sich mit einer KI zu unterhalten.“ (P02)

„Gespräche mit KI fühlen sich für mich zum Start immer etwas komisch an, was sich bessert, umso länger man mit der KI spricht.“ (P03)

Andere hoben hervor, dass der IVA aufgrund seiner Mimik und Gestik künstlich oder störend wirkte:

„Er hat so gestarrt.“ (P01)

„Interaktion mit KI-Agenten [...] fühlt sich aufgrund von Mimik und Gestik auf dem technischen Stand von heute einfach nicht gut an.“ (P27)

Daneben nannten einige Teilnehmende auch **positive Erfahrungen**, etwa eine unerwartet angenehme Gesprächsatmosphäre oder den spielerischen Charakter der Interaktion:

„Ich fand es lustig, mit dem KI-Agenten zu interagieren [...]“ (P25)
„[...] außerordentlich überrascht [...] ob des präzisen Eingehens des KI-Agenten auf meine Antworten [...]“ (P14)

„Nach kurzer Eingewöhnung sehr angenehme Form der Befragung. Ich bin davon sehr überzeugt.“ (P20)

„Es war nett, mit dem Bot zu sprechen, er hat mir kurzes Feedback gegeben, wie er meine Antwort verstanden hat. Es war versichernd zu hören, dass das System meine Äußerung in anderen und manchmal auch besseren Worten wiedergegeben hat.“ (P36)

Andere Kommentare verwiesen schließlich auf **technische Aspekte und Reaktionsverhalten** wie Verzögerungen oder Wiederholungen in den Antworten, die das Erleben beeinträchtigten:

„Die ständige Wiederholung [...] hat mich gestört.“ (P09)

„Er war mir etwas zu langsam und ich fand ihn überfreundlich.“ (P17)

6.7.2 Vertraulichkeit der Antworten

Bezüglich der empfundenen Vertraulichkeit (z. B. Datenschutz oder Privatsphäre) der Antworten gaben 44 % der Teilnehmenden an, sich sicher oder sehr sicher gefühlt zu haben, während 36 % neutral antworteten und 20 % eher oder sehr unsicher waren (vgl. Abb. 6.9).

In den offenen Kommentaren ließen sich drei häufig genannte Einflussfaktoren erkennen. Ein Teil der Teilnehmenden verwies auf die **Mithörbarkeit in der realen Umgebung** und berichtete, beim Sprechen in VR das Gefühl zu haben, von außenstehenden Personen belauscht werden zu können:

„Die Fragen und Antworten werden gesprochen, daher kann jeder in Hörweite verstehen, was passiert.“ (P18)

„Beeinflusst hat auf jeden Fall, dass Menschen in der echten Welt meine Antworten trotzdem hören können, wenn ich sie laut sage.“ (P14)

Darüber hinaus äußerten mehrere Teilnehmende **grundsätzliche Skepsis gegenüber der Datensicherheit von KI-Systemen**, etwa durch die Möglichkeit externer Verarbeitung oder Nutzung für Trainingszwecke:

„Wie soll ich einer KI vertrauen?“ (P02)

„Ich weiß nicht, ob das Sprachmodell lokal lief oder auf irgendwelchen Servern irgendwo und daher die Daten verarbeiten könnte.“ (P28)

„Man weiß nicht, was für Mechanismen der Datenauswertung neben dem eigentlichen Fragebogen noch vom Chatbot genutzt werden.“ (P36)

Schließlich gab es Teilnehmende, die explizit ein **Gefühl von Privatsphäre** beschrieben oder die Vertraulichkeit als **weniger relevant** einstufen, indem sie auf den Studienkontext verwiesen oder angaben, generell unbesorgt zu sein:

„[...] Ich wusste jedoch, dass diese in der Testumgebung niemanden Externen interessieren.“ (P14)

„Ich hab nicht wirklich daran gedacht, wie vertraulich meine Antworten behandelt werden, da mir für diese Antworten die Vertraulichkeit ziemlich egal war.“ (P25)

6.7.3 Vertrauenswürdigkeit des IVA

Die Einschätzungen zur Vertrauenswürdigkeit des IVA fielen gemischt aus. Etwa die Hälfte der Teilnehmenden bewertete ihn als eher oder sehr vertrauenswürdig (11 % „sehr“ und 36 % „eher“), 39 % antworteten neutral, während 14 % den Agenten als eher oder überhaupt nicht vertrauenswürdig einstuft (vgl. Abb. 6.9).

In den offenen Kommentaren ließen sich mehrere Dimensionen erkennen, die das Vertrauen in den IVA prägten. Zum einen wurde die **technologische Vertrauensbasis** thematisiert. Mehrere Teilnehmende äußerten ein generelles Misstrauen gegenüber digitalen Systemen oder KI:

„Generelles Misstrauen gegenüber allem, was digital passiert.“ (P02)

„Es ist ungewohnt, mit einem künstlichen Wesen in freier Sprache zu kommunizieren. Dass dahinter ein Programm und eine Maschine stehen, schafft eher Skepsis.“ (P18)

Auch die **soziale Glaubwürdigkeit** des Agenten, die sich aus Aussehen, Stimme, Intonation und Bewegung ableitete, spielte eine Rolle. Während der Agent einerseits als entspannt beschrieben wurde, wurde andererseits eine übermäßig freundliche Ausdrucksweise oder unnatürliche Wirkung erwähnt:

„Zu freundlich und zu mechanisch [...]“ (P17)

„Wirkte unecht, unsympathisch, Mundbewegungen nicht gut synchronisiert, Körperbewegung sehr abgehackt und irritierend, man möchte am liebsten wegschauen.“ (P11)

„Gut angezogen, entspannte und verständnisvolle Intonation. Aber ein bisschen Uncanny Valley manchmal, vor allem bei Haut und Mimik.“ (P14)

6.7.4 Verständlichkeit der Fragen

Die Verständlichkeit der Fragen wurde von der großen Mehrheit als hoch eingeschätzt. Bei den Likert-Skala-Fragen bewerteten 78 % die Verständlichkeit als sehr oder eher leicht und 8 % als neutral, während 14 % angaben, dass es ihnen eher schwer fiel, die Fragen zu verstehen (vgl. Abb. 6.9). Für die Freitextfragen fiel die Einschätzung noch positiver aus: 94 % empfanden diese als sehr oder eher leicht verständlich und 6 % als neutral.

In den offenen Kommentaren betonten viele Teilnehmende die gute Verständlichkeit, häufig unterstützt durch die **Kombination aus gesprochener und schriftlicher Präsentation**:

„Die Stimme des Agenten war klar und deutlich und außerdem wurden die Fragen weiterhin schriftlich präsentiert.“ (P25)

„Die Frage wurde ja zusätzlich auch noch visualisiert.“ (P18)

Einige wenige berichteten dagegen von Irritation durch die **künstlich wirkende Sprachdarstellung** („Die unechte Ansprache ist irritierend [...]“, P11) oder empfanden insbesondere die Likert-Skalen als belastend („Bei Skalenfragen ist der Agent störend und stresst [...]“, P15). Außerdem wurde, mit Bezug auf die Likert-Fragen, zweimal geäußert, dass „die verneinten Formulierungen etwas schwierig zu verstehen“ (P33) sind.

6.7.5 Beantworten von Likert-Fragen

Die große Mehrheit der Teilnehmenden bewertete die Fähigkeit des IVA, ihre Antworten auf Likert-Fragen korrekt zu erfassen, als gut oder sehr gut (81 %). Etwa 6 % antworteten neutral, während rund 14 % die Erkennung als eher schlecht einschätzten (vgl. Abb. 6.9).

Die qualitativen Kommentare verdeutlichen, dass die Erfahrungen stark variierten. Ein Teil der Teilnehmenden berichtete von einer **sehr zuverlässigen Erfassung**, die in einigen Fällen sogar positiv überraschte:

„Ich war deutlich überrascht, dass der Agent präzise und klar auf vage Aussagen von mir eingegangen ist und vor der ‚Wertung‘ meiner Antwort zunächst sowohl Rückfragen gestellt hat, als auch deutlich auf die Besonderheit meiner vagen Antwort eingegangen ist, diese sogar mit Kontext gefüttert hat.“ (P10)

„Als ich einmal eine Skala mit zwei verschiedenen Werten beantwortete, hat er nachgehakt. Da fühlte ich mich verstanden.“ (P33)

Andere kritisierten hingegen **Verzögerungen**, etwa durch die hohe Latenz, und **ständige Wiederholungen**. Auch ein erhöhter mentaler Aufwand wurde kritisiert:

„Ich finde es zumindest anstrengender, über die Zahlen 1 – 5 im Kopf nachzudenken, als eine Skala vor mir zu sehen und die Antwort so visuell abzuschätzen.“ (P28)

Einmal wurde von einer **Fehlinterpretation** berichtet:

„Es wurde ein Mal ‚zwei‘ statt ‚drei‘ verstanden.“ (P04)

6.7.6 Beantworten von Freitextfragen

Die Formulierung eigener Antworten fiel der Mehrheit der Teilnehmenden leicht: 89 % bewerteten das Formulieren als sehr oder eher leicht, etwa 8 % antworteten neutral und nur eine Person gab an, dass es ihr eher schwer fiel (vgl. Abb. 6.9).

In den offenen Kommentaren betonten viele Teilnehmende die **gute Erfassung und Verarbeitung ihrer Antworten**:

„Die Zusammenfassung meines Gesagten war 100 % zutreffend.“ (P01)

„Die Antworten hat er erstaunlich gut verstanden und – sogar mit anderen Worten – wiederholt.“ (P17)

Die Zusammenfassung der eigenen Antworten wurde teilweise aber auch als störend empfunden.

„[...] das Warten auf die Wiederholung war nervig.“ (P23)

Darüber hinaus beschrieben mehrere Teilnehmende Einschränkungen durch den **Gesprächsfluss des Agenten**, der das Gefühl vermittelte, dass Pausen nicht möglich seien:

„Freitextantworten werden während kurzer Pausen optisch unterbrochen, weil die Person [...] eine Denkpose einnimmt und man das Gefühl hat, dass sie das Gesagte jetzt verarbeitet und ggf. nichts hinzugefügt werden kann.“ (P23)

„[...] Deshalb war man etwas gezwungen, den Redefluss nicht zu unterbrechen.“ (P14)

Einzelne Teilnehmende gaben zudem an, dass die mündliche Formulierung für sie ungewohnt oder mit zusätzlichem Druck verbunden war:

„Ich bin es nicht gewohnt formal und korrekt zu sprechen und fühle mich dadurch unter Druck gesetzt. Schreiben ist mir da lieber.“ (P28)

6.7.7 Auftreten von Schwierigkeiten

Die Hälfte der Teilnehmenden berichtete, während der Befragung Schwierigkeiten erlebt zu haben. Am häufigsten wurden **technische Probleme** (28 %) oder **Probleme mit dem Gesprächsfluss** (22 %) genannt, während **Verständnisprobleme** (14 %) nur vereinzelt auftraten (siehe Tab. 6.15).

Tabelle 6.15: Auftreten von Schwierigkeiten während der Befragung

Kategorie	n	%	
Technische Probleme	10	27,8	██████████
Probleme mit Gesprächsfluss	8	22,2	██████████
Verständnisprobleme	5	13,9	██████
Keine Probleme	18	50,0	████████████████████

Anmerkung: N = 36; Mehrfachantworten waren möglich

Zu den **technischen Problemen** gehörten unter anderem Aussetzer im Mikrofon („Am Anfang hat das Mikro nicht richtig funktioniert und musste neu gestartet werden.“, P25), Störungen durch Umgebungsgeräusche („Durch Umgebungsgeräusche wurde die Übermittlung gestört.“, P08) sowie Darstellungsfehler in der UI. Vereinzelt wurde zudem ein Wechsel der Aussprache beschrieben, bei dem deutsche Sätze mit englischer Intonation wiedergegeben wurden („Der KI-Agent hatte einen deutschen Satz plötzlich mit englischer Aussprache gesprochen, war aber immer noch verständlich.“, P03).

Probleme mit dem Gesprächsfluss traten insbesondere dann auf, wenn Teilnehmende den Agenten unterbrachen oder Pausen einlegten. In diesen Situationen entstand das Gefühl, dass der Gesprächsverlauf gestört wurde:

„Außerdem hatte ich das Gefühl, den Gesprächsfluss des Agenten komplett durcheinanderzubringen, wenn ich außerhalb der Antwortfenster etwas gesagt habe.“ (P25)

Zudem wurde das Fehlen einer Füllwörter-Erkennung als störend empfunden:

„Manchmal hatte der Avatar schon seine Animation gewechselt, als ich nur ein 'hm' gesagt hatte. Da dachte ich, dass ich mich mit meiner Antwort beeilen muss, [...] obwohl ich gerne noch etwas länger überlegt hätte.“ (P36)

Mehrere Teilnehmende beschrieben außerdem das Gefühl, dass die KI Antworten zu früh auswertete, oder fühlten sich unterbrochen:

„[...] hatte ich zwischendurch das Gefühl, der Agent beginnt ‚zu früh‘ mit der Wortauswertung.“ (P22)

„Bei den Freitextantworten fühlte ich mich manchmal unterbrochen.“ (P23)

Auch lange Antwortlatenzen wirkten sich störend auf den Gesprächsfluss aus, vor allem bei Likert-Skalen:

„Hat manchmal lange gedauert, um eine einzelne Zahl zu verstehen.“ (P02)

Verständnisprobleme spielten insgesamt eine geringe Rolle. Ein Teilnehmender berichtete von anfänglicher Verwirrung über die Art der Eingabe:

„Missverständlich, wusste nicht, ob ich irgendetwas anklicken sollte, habe erst nach einigen Fehlversuchen verstanden, dass ich sprachlich antworten sollte.“ (P11)

Ein Vorfall betraf eine zunächst fehlerhafte Zuordnung einer Antwort: Ein Teilnehmender berichtete, der Agent habe *„erst nicht reagiert und danach doppelt“* (P15), sodass die Eingabe der nachfolgenden Frage zugeschrieben wurde. Der oder die Teilnehmende machte die Studienleitung auf das Problem aufmerksam und erhielt die Anweisung, den IVA aufzufordern, die Frage zu wiederholen. So konnte die Frage erneut beantwortet und korrekt erfasst werden.

Insgesamt zeigen diese Angaben, dass die meisten Probleme zwar nur vereinzelt und situativ auftraten, sie jedoch den Gesprächsfluss gelegentlich deutlich beeinträchtigen konnten.

6.7.8 Präferenzen zu Erscheinungsbild und Aussehen

Bei der Frage nach dem bevorzugten Erscheinungsbild des IVA zeigte sich, dass nur wenige Teilnehmende ein bestimmtes Geschlecht bevorzugten (männlich: 2,8 %, weiblich: 13,9 %, neutral/androgyn: 13,9 %). Die Mehrheit wünschte sich stattdessen eine abstrakte oder roboterartige Darstellung (36,1 %), während 27,8 % keine Präferenz angab. Zwei Teilnehmende (5,6 %) sprachen sich für eine körperlose Darstellung aus (vgl. Tab. 6.16).

Tabelle 6.16: Präferenzen für das Erscheinungsbild des KI-Agenten ($n = 36$)

Kategorie	n	%	
Männlich	1	2,8	■
Weiblich	5	13,9	■■■■■
Neutral / androgyn	5	13,9	■■■■■
Abstrakt / roboterartig	13	36,1	■■■■■■■■■■■■■■■
Körperlos	2	5,6	■■
Keine Präferenz	10	27,8	■■■■■■■■■■

In den offenen Kommentaren zum Aussehen, Verhalten und zur Stimme des Agenten äußerten viele Teilnehmende den Wunsch nach einer **abstrakten oder nicht-menschlichen Gestaltung**, um Irritationen durch das *Uncanny Valley* (vgl. Abschnitt 2.2.1 [Darstellung und Verhalten](#)) zu vermeiden. Besonders die eher unrealistische Mimik und die Mundbewegungen wurden wiederholt als befremdlich beschrieben:

„He definitely sat in the 'uncanny valley' [...] especially with the moving mouth. However, I thought about this less as time went on. I think a more abstract agent

would fit better.“ (P13)

„Der KI Agent sah noch recht künstlich aus, gerade die Mimik im Gesicht. Eine abstrakte Gestalt würde mir besser gefallen.“ (P05)

Mehrere Teilnehmende betonten zudem, dass eine weniger menschliche, roboterartige oder stilisierte Figur helfen könne, problematische Zuschreibungen (z. B. Geschlecht, Ethnie, soziale Zugehörigkeit) zu vermeiden:

„Je abstrakter, desto höher die Akzeptanz, würde ich sagen.“ (P15)

„Norschön [sic] weißer Mann wäre nicht meine erste Wahl für einen KI-Agenten. Abstrakt / roboterartig wäre besser und angenehmer.“ (P23)

Manche empfänden eine abstraktere Gestaltung auch als „ehrlicher“, da sie die Künstlichkeit nicht zu verschleiern versuche:

„Ich fände eine abstraktere Darstellung besser, da sie die KI besser darstellen würde und somit ehrlicher wäre. Zu versuchen, wie ein realistischer Mensch aufzutreten, wirkt bei so etwas wie ein Täuschungsversuch auf mich.“ (P28).

*„Weniger menschliche Mimik, die man als User*in immer zu deuten versucht.“* (P23)

Andere Stimmen machten deutlich, dass das **Aussehen insgesamt zweitrangig** sei, solange Funktionalität, Stimme und Verhalten stimmten. So beschrieben Teilnehmende den Agenten eher als *Werkzeug* oder *Lautsprecher*:

„Das ist mir persönlich egal, ich würde mich auch mit einem digitalen Mülleimer unterhalten.“ (P02)

„Das Aussehen spielte für mich keine große Rolle, da ich das Gefühl hatte, die Antworten eher der Tafel mit den geschriebenen Fragen zu geben und nicht dem KI-Agenten, der für mich eher die Funktion eines Lautsprechers übernommen hat.“ (P09)

Wiederholt wurde auch die **Kontextabhängigkeit der Gestaltung** betont. Einige hielten es für sinnvoll, die visuelle Gestaltung stärker auf spezifische Einsatzbereiche oder Umgebungen abzustimmen. Andere regten an, das Erscheinungsbild an die jeweilige Situation oder sogar an individuelle Präferenzen anzupassen:

„Je nach Kontext würde ich eine anderes Aussehen bevorzugen. [...] Am besten wäre es natürlich, die Präferenz der zu befragenden Person vorab zu kennen und dementsprechend den Agenten setzen zu können.“ (P10)

Neben den kritischen Rückmeldungen gab es auch **positive Bewertungen des Erscheinungsbildes und Verhaltens**. Mehrere Teilnehmende beschrieben den Agenten als angenehm oder neutral:

„Aussehen und Verhalten waren neutral und ruhig, was nicht so stresst.“ (P06)
 „Das Aussehen des Agenten wirkte auf mich wie eine seriöse virtuelle Persönlichkeit.“ (P29)
 „Optisch habe ich die Person kaum wahrgenommen, die Stimme war klar, verständlich und das Verhalten höflich.“ (P08)

Manche Teilnehmende erwähnten, dass der Agent mit besserer Darstellung und Verhalten einen positiveren Eindruck erwecken würde:

„Der KI-Agent wäre cooler, wenn er nicht so unsympathisch starren würde.“ (P01)
 „Bessere Gestik und Motorik, damit das Gegenüber noch natürlicher wirkt.“ (P18)

Insgesamt deuten die Ergebnisse darauf hin, dass eine klare Mehrheit eine **abstrakte oder roboterartige Darstellung** bevorzugt. Eine neutrale oder flexible Gestaltung könne am ehesten Akzeptanz finden und gleichzeitig Irritationen durch das Uncanny Valley vermeiden.

6.7.9 Präferenzen für Antwortformate

Die Präferenzen für die beiden Antwortformate – Sprachinteraktion mit dem IVA vs. Texteingabe über die UI – variierten deutlich in Abhängigkeit vom Fragetyp (siehe Tab. 6.17, Abb. 6.10). Zur Prüfung, ob die Ergebnisse signifikant von einer zufälligen Verteilung abweichen, wurde ein Einstichproben-z-Test durchgeführt (vgl. Abschnitt 4.5). Ein signifikanter p -Wert ($p < ,05$) weist auf eine ungleich verteilte Präferenz hin.

Tabelle 6.17: Präferenzen für IVA vs. UI in Abhängigkeit vom Fragetyp

Fragetyp	n_{IVA}	$p_{IVA}(\%)$	n_{UI}	$p_{UI}(\%)$	z	p
Likert-Skalen	10	27,8	26	72,2	-2,98	,003
Freitextfragen	27	75,0	9	25,0	3,46	<,001
Gesamtpräferenz	23	63,9	13	36,1	1,74	,083

Anmerkung: $n = 36$; z-Test gegen gleichverteilte Präferenz ($p_0 = .50$).

Für die Beantwortung von **Likert-Skalen** wurde die UI häufiger bevorzugt (72,2 % vs. 27,8 %), ein Unterschied, der auch statistisch signifikant war ($z = -2,98$, $p = ,003$). Bei den **Freitextfragen** zeigte sich das entgegengesetzte Muster: Eine klare Mehrheit bevorzugte den IVA (75,0 % vs. 25,0 %), ebenfalls mit einem signifikanten Unterschied ($z = 3,46$, $p < ,001$). Betrachtet man die **Gesamtpräferenz** über beide Fragetypen hinweg, so gaben 63,9 % der Teilnehmenden an, den IVA zu bevorzugen, während 36,1 % die UI bevorzugten. Dieser Unterschied war statistisch nicht signifikant ($z = 1,74$, $p = ,083$).

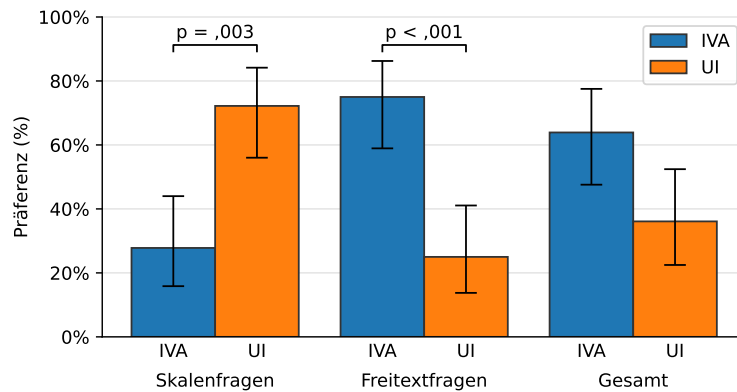


Abbildung 6.10: Präferenzen der Teilnehmenden für IVA vs. UI in Abhängigkeit vom Fragetyp (prozentuale Anteile mit 95%-Konfidenzintervallen, $n = 36$)

Präferenz bei Likert-Skalen

Für die Beantwortung der Likert-Skalen wurde die klassische Texteingabe in der UI deutlich häufiger bevorzugt. In den Freitextkommentaren nannten viele Teilnehmende als Hauptgründe für die **UI-Präferenz** die größere **Schnelligkeit und Einfachheit** der Eingabe. Das direkte Auswählen einer Zahl wurde als „*einfacher, schneller*“ (P10) und besser kontrollierbar beschrieben:

„Die direkte Auswahl von Zahlen geht besser von der Hand.“ (P05)

„[...] weil es dort keine Latenz gibt [...].“ (P36)

Zudem bot die UI den Vorteil, **ohne Zeitdruck** zu antworten und sich bei Bedarf umentscheiden zu können:

„So habe ich nicht das Gefühl, dass jemand auf meine Antwort wartet.“ (P08)

Auch die **visuelle Abschätzung der Skala** wurde als hilfreich empfunden:

„[...] kann auch schauen, ob der optische Eindruck der markierten Skala zu meiner Einschätzung passt.“ (P14)

Für die IVA-Variante nannten einzelne Teilnehmende vor allem den höheren **Spaß- und Interaktionsfaktor**:

„Das macht auf jeden Fall mehr Spaß, als irgendwelche Haken oder Markierungen zu setzen.“ (P19).

Es wurde auch betont, dass die Sprachinteraktion eine intuitive und natürliche Alternative darstelle, insbesondere wenn es um Barrierefreiheit gehe:

„[...] dass ich den Sinn von der sprachbasierten Eingabe in diesem Fall nur sehe, wenn man seine Hände währenddessen für andere Dinge benötigt oder generell nicht benutzen kann.“ (P25)

„I do see that it solves some issues regarding accessibility in VR though.“ (P16)

Einige Teilnehmende wiesen darauf hin, dass die UI bei Likert-Skalen weiterhin Vorteile biete, auch wenn der IVA technisch überzeugend sei:

„KI-Agent, aber nur, wenn gut umgesetzt und keine Skalenfragen.“ (P15)

Präferenz bei Freitextfragen

Für die Beantwortung der Freitextfragen wurde die Sprachinteraktion mit dem IVA von der großen Mehrheit bevorzugt. Als Hauptgründe nannten die Teilnehmenden die **geringere Anstrengung** und den **natürlicheren Gesprächsfluss** beim Sprechen. Das Formulieren über Sprache wurde als schneller, intuitiver und weniger mühsam erlebt als die Texteingabe über die VR-Tastatur:

„Es ist wesentlich anstrengender, in VR einen Text manuell zu verfassen, als einfach zu sprechen und die KI wertet das aus.“ (P02)

„Die Spracheingabe ist sehr viel angenehmer für Texte und hat gut funktioniert.“ (P14)

Viele beschrieben die Sprachinteraktion als **stressfreier und flüssiger** und betonten, dass sie dadurch ausführlicher und spontaner antworten können:

„It is so much easier than typing, and as a result I can be much more detailed.“ (P13)

Darüber hinaus wurde die IVA-Befragung als **komfortabler, weniger belastend und unterhaltsamer** erlebt:

„Insgesamt macht die Interaktion mit dem KI Agent mehr Spaß und ist für längere Eingaben einfacher zu bedienen.“ (P05)

„Die Bedienung ist sehr intuitiv und braucht auch nicht noch die Kenntnis, mit einem VR-Kontroller umgehen zu können.“ (P08)

„It feels less like I'm filling out some bureaucratic form, or completing an exam.“ (P13)

Einzelne Teilnehmende wiesen jedoch auf **Einschränkungen der Sprachvariante** hin. So wurde bemängelt, dass es keine Möglichkeit gebe, Antworten nachträglich zu korrigieren oder zu überarbeiten:

„[Es] fehlt mir allerdings, dass ich meine Antwort Korrekturlesen kann, um sicherzugehen, dass meine Antwort auch genau so abgespeichert wird, wie ich sie formuliert habe.“ (P23)

Andere äußerten den Wunsch nach einer Kombination beider Ansätze:

„Ich selbst würde eine hybride Befragung bevorzugen. [...] Ein digitales Formular mit Skalenfragen. Dann abschließend [...] eine Befragung des Agenten für Freitextfragen.“ (P10)

„Oder eine Kombination aus beiden, wo ich sowohl manuell was eingeben kann, als auch reden kann.“ (P24)

Wenige Teilnehmende gaben schließlich an, die UI auch bei Freitextfragen als neutraler oder kontrollierbarer zu empfinden oder den Agenten als ablenkend wahrzunehmen:

„Ich bin gefühlt weniger unter Druck, wenn nicht jemand (anscheinend) auf meine Antwort wartet, sondern ich (gefühlte) das Tempo selbst bestimmen und die Antwort einfach und selbständig korrigieren kann.“ (P17)

„Der KI-Agent ist unnötig für eine einfache Umfrage und stört nur.“ (P35)

Gesamtpräferenz für die Befragungsart

Für beide Fragetypen hinweg gaben 63,9 % der Teilnehmenden an, die Befragung mit dem IVA zu bevorzugen, während 36,1 % die klassische Texteingabe in der UI bevorzugten. Dieser Unterschied war statistisch nicht signifikant. Insgesamt deuten die qualitativen Befunde konsistent darauf hin, dass die UI vor allem bei Likert-Skalen Vorteile bietet, während die Sprachinteraktion besonders bei Freitextfragen als angenehmer und passender erlebt wird.

6.7.10 Verbesserungsvorschläge

In den abschließenden Kommentaren machten die Teilnehmenden vor allem konkrete Vorschläge zur Weiterentwicklung des IVA.

Ein zentrales Thema betraf den **Gesprächsablauf**. Mehrere Teilnehmende regten an, Unterbrechungen und Wiederholungen – sowohl in den Agenten-Äußerungen als auch bei der Bestätigung von Antworten – zu reduzieren:

„Die ständige Wiederholung der immer gleichen Antwortmöglichkeiten hat mich gestört.“ (P09)

„Weniger dazwischenreden, wenn nicht direkt eine Antwort auf eine Frage kommt.“ (P02)

„Das war etwas nervig [...], dass der Agent immer die Antwort wiederholt hat. Ich hätte mir eher ein visuelles Feedback am UI gewünscht.“ (P24)

Zudem wurde vorgeschlagen, eine Möglichkeit einzuführen, Antworten vor dem Fortfahren zu bestätigen oder bei Bedarf zu korrigieren:

„The AI doesn't give you a chance to correct your answer before submitting, or introduce the ability to confirm your answer or correct it before moving on.“ (P16)

Schließlich wurde auch der Wunsch nach einer natürlicheren Gesprächsführung geäußert:

„Die einfache zwischenmenschliche Interaktion mit dem Agenten fehlt bzw. wurde von mir gar nicht für möglich gehalten.“ (P34)

Darüber hinaus wurden auch **technische Aspekte** thematisiert. Genannt wurden insbesondere kürzere Latenzen und eine natürlichere Gestik:

„Interaktion ist noch sehr künstlich und die Reaktion des KI-Agenten verzögert, eher langsam, das könnte verbessert werden. Allerdings ist es erstaunlich, was überhaupt möglich ist an Interaktion mit einem solchen Agenten.“ (P17)

Mit Blick auf die **Gestaltung des Studienablaufs** schlugen Teilnehmende vor, den Agenten flexibler einzusetzen:

„Agent entweder erst erscheinen lassen, wenn er gebraucht wird oder Interaktion schon vorher ermöglichen.“ (P15)

Zudem wurde die gleichzeitige Anzeige des Fragebogens kritisch bewertet:

„Der KI-Agent müsste im Mittelpunkt stehen oder sitzen und der Fragebogen gar nicht angezeigt werden. Dadurch würde man sich stärker auf den Agenten konzentrieren. [...] Befragung und selber Lesen der Frage [...] ist zu viel und wird bei einer Befragung durch eine menschliche Person eigentlich so auch nicht gemacht.“ (P34)

Zudem wurden erneut **hybride Befragungssysteme** angeregt, die Spracheingabe und klassische UI-Elemente kombinieren.

Trotz der genannten Kritikpunkte hoben mehrere Teilnehmende hervor, dass die KI insgesamt einen „guten Job“ gemacht habe und die Befragung als neuartig, interessant und angenehm empfunden wurde:

„Es hat auf jeden Fall zusätzlichen Unterhaltungswert und die Neuartigkeit ist auch interessant.“ (P14)

6.8 Technische Kennzahlen und Leistungsanalyse

In diesem Abschnitt werden die während der Studiendurchläufe erfassten Systemkennzahlen zusammengefasst. Sie beinhalten die Latenzen der Sprachpipeline, das Interaktionsverhalten der Nutzenden sowie die Erkennungs- und Fehlerraten der einzelnen Systemkomponenten.

6.8.1 Latenzen der Sprachpipeline

Die folgenden Analysen beschreiben die gemessenen Latenzen der Sprachpipeline, also die zeitlichen Verzögerungen zwischen Nutzeräußerung, Modellverarbeitung und Wiedergabe durch den Agenten.

Abb. 6.11 stellt die Latenzen der einzelnen Verarbeitungsschritte im Vergleich dar. Besonders deutlich wird, dass die größte Verzögerung durch die LLM-Verarbeitung entsteht, während ASR und TTS weitgehend im akzeptablen Bereich bleiben. Die Playback-Latenz, also das Delay beim Übertrag von Python zur Wiedergabe in Unity über WebSocket, fiel im Mittel sehr gering aus und ist im Vergleich zu den anderen Komponenten vernachlässigbar.

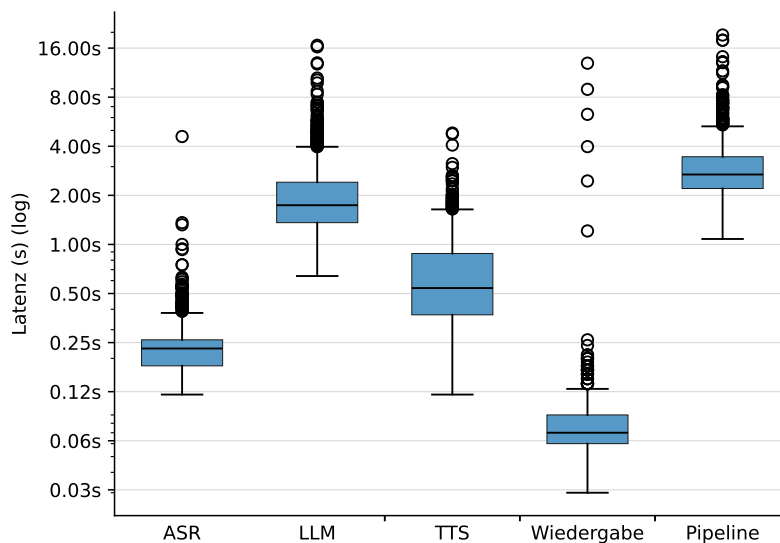


Abbildung 6.11: Vergleich der Latenzen über die Verarbeitungsstufen hinweg (log-Skala)

Wie Tab. 6.18 zeigt, lag die durchschnittliche Gesamtlatenz der Sprachpipeline, also die Zeitspanne vom Ende einer Nutzeräußerung bis zum Beginn der IVA-Wiedergabe, bei rund drei Sekunden ($M = 3,11$ s, $SD = 1,67$, Median = 2,69 s). Damit überschritt die Pipeline weiterhin die in der Literatur empfohlene Zwei-Sekunden-Grenze für flüssige Konversationssysteme. Der größte Anteil an der Verzögerung entfiel mit rund 69 % auf die Verarbeitung durch das Sprachmodell (LLM), das im Mittel 2,16 Sekunden benötigte.

Tabelle 6.18: Latenzen der Sprachpipeline (in Sekunden, n variiert je nach Komponente)

Komponente	n	M	SD	Median	% von Gesamt
ASR	1085	0,24	0,16	0,23	7,8
LLM	1014	2,16	1,44	1,74	69,3
TTS	970	0,68	0,49	0,54	21,9
Übertragung	942	0,12	0,57	0,07	3,7
Gesamt (Pipeline)	942	3,11	1,67	2,69	100,0

Anmerkung: Gesamt (Pipeline) umfasst nur abgespielte Gesprächsrunden, die einzelnen Komponenten auch unterbrochene Runden

Die ASR und die TTS waren mit durchschnittlich 0,24 bzw. 0,68 Sekunden deutlich schneller und trugen zusammen weniger als ein Drittel zur Gesamtlatenz bei. Die Übertragung der Audiodaten von Python zu Unity über WebSocket spielte im Mittel nur 0,12 Sekunden eine vernachlässigbare Rolle.

Abb. 6.12 verdeutlicht die Verteilung der Gesamtpipeline-Latenzen: Die meisten Interaktionen lagen im Bereich zwischen zwei und vier Sekunden, gleichzeitig traten aber auch längere Verzögerungen von bis zu 19,3 Sekunden auf. Die rechtsschiefe Form der Verteilung weist damit auf einzelne, aber deutliche Ausreißer hin.

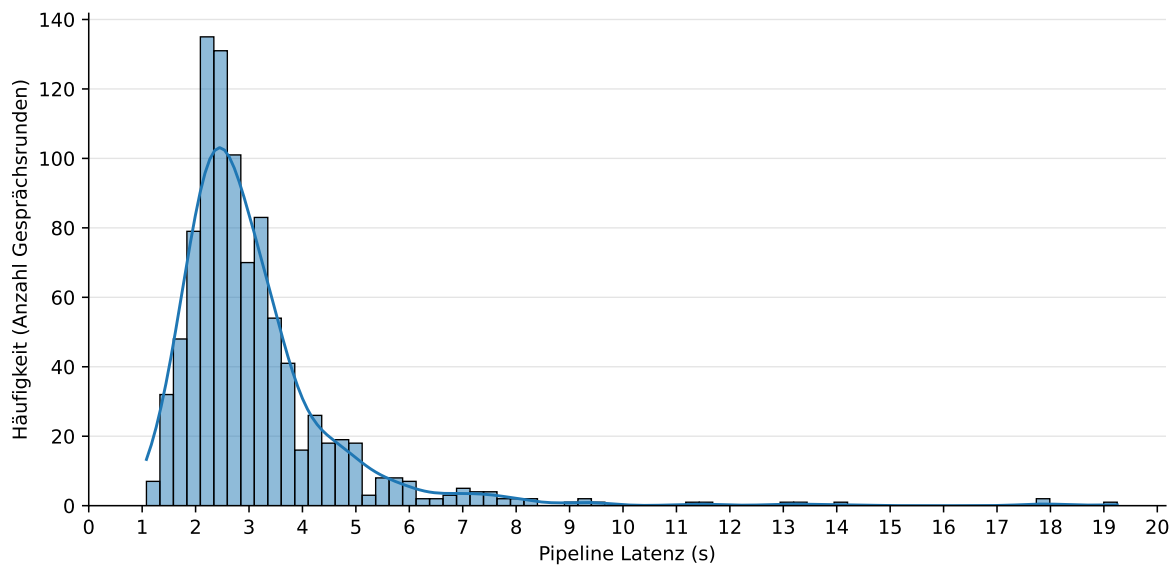


Abbildung 6.12: Histogramm der Gesamt-Pipeline-Latenzen (Die Hälfte der Werte liegt zwischen 2,2 s und 3,5 s, mit vereinzelt Ausreißern bis 19,3 Sekunden)

6.8.2 Nutzerverhalten und Interaktionsdynamik

Neben den Systemlatenzen wurde auch das Verhalten der Nutzenden während der Gespräche analysiert. Insgesamt wurden 1159 Gesprächsrunden mit 36 Teilnehmenden protokolliert, davon wurden 923 zumindest teilweise durch den IVA wiedergegeben. Eine Gesprächsrunde wurde definiert als ein Durchlauf der Sprachpipeline, angefangen mit der Spracherkennung der Teilnehmenden und endend in der Wiedergabe des Agententextes. Eine Gesprächsrunde konnte aber auch durch einen Prompt ans LLM vom System zu Beginn des Gesprächs ausgelöst werden. Oder während des Gesprächs, wenn die oder der Nutzer:in nach zehn Sekunden noch keine Antwort gegeben hatte. In 783 dieser Runden hatte das LLM eines der Fragebogen-Items geäußert und in 770 Runden hatten die Teilnehmenden eine Antwort auf ein Item gegeben.

Neben den Systemlatenzen wurde auch die *User-Latenz* erfasst, also die Zeitspanne zwischen dem Ende der Agentenantwort und dem Beginn der nächsten Nutzereingabe. Abb. 6.13 zeigt die Verteilung dieser Latenzen: Negative Werte kennzeichnen Unterbrechungen, also Fälle, in denen die Nutzenden bereits sprachen, bevor der Agent seine Antwort vollständig beendet hatte. Das Histogramm zeigt eine rechtsschiefe Verteilung mit einer hohen Dichte kurzer Reaktionszeiten.

In insgesamt 1032 Gesprächsrunden antworteten die Teilnehmenden dem Agenten. Davon entfielen 304 auf solche Unterbrechungen, was einem Anteil von 29,5 % entspricht. Für die verbleibenden 728 Gesprächsrunden lag die mittlere User-Latenz bei $M = 3,59$ Sekunden ($SD = 4,43$ s, $Median = 1,94$ s). Die meisten Interaktionen erfolgten somit innerhalb von ein

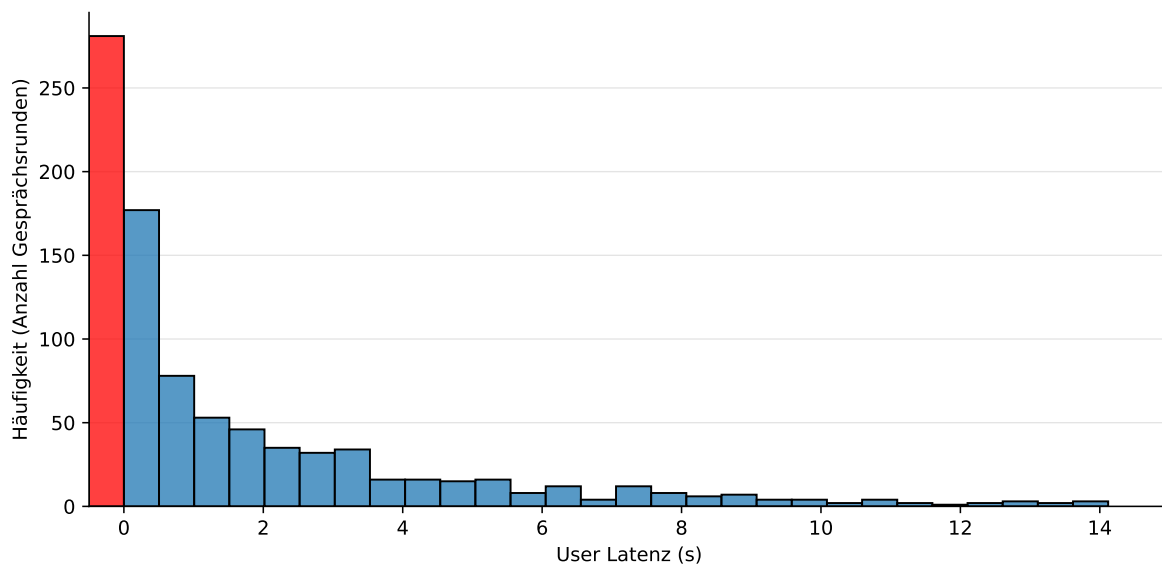


Abbildung 6.13: Verteilung der User-Latenzen (Zeit bis zur nächsten Nutzereingabe), negative Werte (in Rot) entsprechen Unterbrechungen der Wiedergabe durch das Ins-Wort-Fallen

bis drei Sekunden nach Ende der Agentenantwort, vereinzelt traten jedoch deutlich längere Pausen von bis zu 28,61 Sekunden auf.

Darüber hinaus wurde erfasst, wie häufig bereits vorbereitete Agentenantworten verworfen wurden, weil die Nutzenden während der Generierung oder Sprachsynthese erneut zu sprechen begannen. In 217 Fällen (etwa 18,7 % der Prozesse) erfolgte eine Unterbrechung noch vor Beginn der Wiedergabe. In den meisten Fällen lässt sich dies darauf zurückführen, dass die Teilnehmenden während der Äußerung Pausen einlegten (z. B. zum Überlegen), wodurch die Sprachpipeline bereits ausgelöst wurde. Die Fortsetzung der Eingabe führte dazu, dass der zuvor erzeugte Prompt verworfen wurde, bevor die generierte Antwort vollständig verarbeitet oder abgespielt werden konnte.

6.8.3 Fehlerrate der Spracherkennung (ASR)

Da für die Sprachaufnahmen keine manuellen Referenztranskriptionen (*Ground-Truth*) vorlagen, konnte keine klassische *Word Error Rate* berechnet werden. Stattdessen wurde überprüft, inwieweit die ASR-Transkriptionen offensichtlich vom jeweils erwarteten Antwortformat abwichen. Bei Likert-Skalen wurden fehlerhafte Erkennungen wie „Fömpf“ oder unverständliche Äußerungen wie „Koller“ als Fehlklassifikationen gewertet. Bei offenen Freitextfragen wurde erwartet, dass die Transkriptionen sprachlich sinnvolle Sätze oder Phrasen enthalten; Fülllaute („äh“, „hm“) wurden dabei nicht als Fehler betrachtet.

Insgesamt wurden 1049 Sprachtranskriptionen durch das ASR-System verarbeitet. Davon waren 988 Transkriptionen korrekt, was einer Erkennungsrate von 94,2 % entspricht. In 5,5 % der Fälle trat eine fehlerhafte Transkription auf, meist bedingt durch phonetisch ähnliche Wörter oder unklare Aussprache (z. B. „Neun“ statt „Nein“, „höchstlich“ statt „körperlich“). Halluzinationen, also inhaltlich erfundene Transkriptionen, traten in fünf Fällen (0,5 %) auf (z. B. „Untertitelung des ZDF, 2020“ oder „Hm hm hm hm hm hm...“).

36,2 % der fehlerhaften Transkriptionen konnten durch das LLM automatisch korrigiert werden, indem das Modell den intendierten Wert aus dem Kontext ableitete („Siem“ zu „Sieben“). In weiteren 25,9 % der Fälle wurden die Eingaben vom LLM ignoriert. In den übrigen Fällen forderte das Modell meist eine Wiederholung der Antwort an („Entschuldige, ich habe dich nicht ganz verstanden. Bitte antworte mit einer Zahl von eins bis sieben.“).

Das ASR-Modul hatte vor allem Schwierigkeiten mit kurzen Äußerungen, die nur aus einem Wort (36,2 % der Fehler) oder zwei Wörtern (12,1 % der Fehler) bestanden. Insgesamt lag die Rate offensichtlicher ASR-Fehler unter 6 %, was auf eine insgesamt robuste Erkennungsleistung der Sprachpipeline hinweist.

6.8.4 Erkennung von Sprachaktivität (VAD)

Für die VAD lagen keine objektiven Messdaten zur Erkennungsgenauigkeit vor. Beobachtungen während der Datenerhebung deuten jedoch darauf hin, dass Sprachaktivität in der Regel zuverlässig erkannt wurde, solange die Teilnehmenden in normaler Lautstärke sprachen. Fehlaktivierungen durch Geräusche wie Räuspfern traten nur vereinzelt auf, und auch Hintergrundgeräusche wurden meist korrekt ignoriert. Bei sehr leiser oder zögerlicher Aussprache kam es gelegentlich zu ausbleibenden Aktivierungen. In solchen Fällen wiederholten die Teilnehmenden ihre Aussage meist unmittelbar. In den offenen Nutzerrückmeldungen wurden ausbleibende Erkennungen des VAD nicht erwähnt.

6.8.5 Fehleranalyse

Zur Bewertung der technischen Zuverlässigkeit wurden alle im Betrieb des Sprachdialogsystems aufgetretenen Fehlermeldungen systematisch erfasst und kategorisiert (siehe Tab. 6.19). Insgesamt wurden 1159 Gesprächsrunden mit 36 Teilnehmenden aufgezeichnet, von denen 1104 (95,25 %) fehlerfrei verliefen.

Tabelle 6.19: Auftreten von Fehlermeldungen in den Gesprächsrunden ($n = 1159$)

Fehlerart	n	%	Einfluss auf den Studienablauf
LLM: Itemnummer stimmt nicht überein	30	2,59	kein Einfluss
LLM: Item übersprungen	13	1,12	veränderte Itemreihenfolge
LLM: Item nicht gestellt	4	0,35	Eingriff der Studienleitung
LLM: Maximale Tokenzahl überschritten	2	0,17	kein Einfluss
Mikrofon nicht erkannt	2	0,17	Neustart der VR-Verbindung
LLM: UTF-8-Dekodierungsfehler	1	0,09	Pipeline-Neustart
ASR: Laufzeitfehler	1	0,09	Pipeline-Neustart
TTS: Keine Audiodaten empfangen	1	0,09	kein Einfluss
Protokollierungsfehler	1	0,09	kein Einfluss
<i>Kein Fehler</i>	1104	95,25	–

LLM-bezogene Fehler

Die meisten Fehlermeldungen wurden durch das Sprachmodell verursacht, welches insgesamt für etwa 92,7 % der Fälle verantwortlich war. Der weitaus größte Teil dieser Fehlermeldungen hatte jedoch keinen Einfluss auf den Ablauf der Studie. In 30 Fällen (2,6 %) gab das LLM eine falsche Itemnummer zur Nutzerantwort aus (*Itemnummer stimmt nicht überein*). Diese Fehler wurden vom Kontrollalgorithmus automatisch erkannt und korrigiert.

In 13 Fällen (1,1 %) übersprang das LLM eine Frage (*Item übersprungen*). Die Folge war eine kurzzeitig veränderte Itemreihenfolge, die jedoch keinen inhaltlichen Einfluss auf den Fragebogen hatte.

In vier Fällen (0,4 %) stellte das Modell eine Frage nicht (*Item nicht gestellt*), obwohl sie im Prompt als offen markiert war. Nach Hinweis der Studienleitung wurde der Agent von den Teilnehmenden manuell zur Fortsetzung aufgefordert, sodass der Durchlauf vollständig abgeschlossen werden konnte.

Zweimal wurde die *Maximale Tokenzahl* der Modellantwort *überschritten* (0,3 %), verursacht durch sehr lange Nutzerantworten. Die relevanten Daten blieben dennoch in den Protokollen erhalten. Im selben Zusammenhang trat in einem Fall ein *ASR-Laufzeitfehler* (0,1 %) auf, ausgelöst durch eine Überschreitung der maximalen Audiolänge (> 30 s).

Ein einmaliger *UTF-8-Dekodierungsfehler* (0,1 %) infolge einer nicht interpretierbaren Modellausgabe führte zu einem kurzzeitigen Pipeline-Neustart.

Technische Fehler

Auch technische Störungen traten nur vereinzelt auf. In zwei Sitzungen (0,3 %) wurde das Mikrofon zunächst nicht erkannt, was durch einen Neustart der VR-Verbindung behoben werden konnte. Ein einzelner Fehler beim Ergebnis-Logging (0,1 %) wurde nachträglich anhand der Logdaten manuell ergänzt. In einem Fall wurden vom TTS keine Audiodaten empfangen, was vermutlich auf eine Netzwerkstörung zurückzuführen ist.

Manuelle Validierung der numerischen Antworten

Zusätzlich zur automatischen Fehlererfassung wurden alle numerischen Antworten während der Erhebung manuell überprüft, um potenzielle Abweichungen zwischen der intendierten und der vom System registrierten Antwort zu identifizieren. Dabei traten in zwei Fällen fehlerhafte Datenzuordnungen auf:

1. In einem Fall wurde der Wert „2“ statt „3“ vom ASR transkribiert; der Fehler fiel dem oder der Teilnehmenden auf, versuchte zwar nicht, den Fehler zu korrigieren, erwähnte den Fehler aber in dem Abschlussfragebogen.
2. In einem weiteren Fall wurde eine Zahl zweimal geäußert und vom LLM der nachfolgenden Frage zugeordnet. Der oder die Teilnehmende erkannte den Fehler selbstständig und konnte ihn nach Rücksprache mit der Studienleitung korrigieren.

Die einmalige fehlerhafte Transkribierung hatte keinen signifikanten Einfluss auf die Ergebnisse. Der Mittelwert der betroffenen IPQ-Subskala (Spatial Presence) würde sich bei der Korrektur von „2“ auf „3“ um 0.005 ändern ($M_2 = 4,277$, $S_2 = 0,885$; $M_3 = 4,272$, $SD_3 = 0,886$). Insgesamt traten somit keine systematischen Abweichungen in der numerischen Answerfassung auf. In keinem Fall gingen Daten verloren.

6.9 Zusammenhänge und Einflussfaktoren

Zur Untersuchung wechselseitiger Beziehungen zwischen den erhobenen Variablen wurde eine Korrelationsanalyse mit Spearman's ρ (rho) durchgeführt (Spearman, 1904). Dieses nicht-parametrische Verfahren bewertet den Zusammenhang zwischen zwei Variablen anhand ihrer Rangordnung und erfordert keine Normalverteilung der Daten. Der Korrelationskoeffizient ρ kann Werte zwischen -1 (perfekter negativer Zusammenhang) und $+1$ (perfekter positiver Zusammenhang) annehmen. Nach den Konventionen von Cohen (1988) werden Beträge von $|\rho| = 0,10$ als klein, $|\rho| = 0,30$ als mittel und $|\rho| = 0,50$ als stark interpretiert.

In die Analyse gingen sowohl demografische Faktoren (z. B. Alter, Vorerfahrung) als auch psychometrische Skalen (Präsenz, Arbeitsbelastung, Gebrauchstauglichkeit, Vertrauen und Präferenz) ein.

Abb. 6.14 zeigt die Korrelationsmatrix der zentralen Variablen; signifikante Zusammenhänge ($p < ,05$) sind farblich hervorgehoben. Höhere Werte in den Skalen stehen jeweils für stärker ausgeprägte Merkmale: Ein höheres *Gesamtgefühl*, eine höhere *Vertraulichkeit* und *Vertrauenswürdigkeit* spiegeln positivere Empfindungen wider. Die Präferenzvariable wurde so kodiert, dass 0 der UI- und 1 der IVA-Bedingung entspricht. Eine vollständige Übersicht aller berechneten Korrelationen findet sich in Tab. C.2 im Anhang.

Präsenz, Vertrauen und Usability

Zwischen den Variablen des Präsenzepfindens (IPQ) und der Vertrauensbewertung zeigten sich deutliche positive Zusammenhänge. Teilnehmende mit höheren **Präsenzwerten** berichteten zugleich höhere Werte im **Vertrauen in den IVA** ($\rho = .71, p < ,001$) und bewerteten das System tendenziell als benutzerfreundlicher ($\rho = .32, p = ,06$). Auch zwischen den Präsenzwerten der IVA- und der UI-Bedingung bestand ein starker positiver Zusammenhang ($\rho = .86, p < ,001$). Das **Gesamtgefühl** während der Befragung ($\rho = .52, p < ,001$) sowie die **Vertrauenswürdigkeit des IVA** ($\rho = .71, p < ,001$) korrelierten ebenfalls positiv mit dem Präsenzepfinden.

Arbeitsbelastung und Usability

Zwischen der wahrgenommenen **Arbeitsbelastung** (NASA-TLX) und der **System Usability** (SUS) zeigten sich negative Zusammenhänge. In der IVA-Bedingung korrelierte der NASA-TLX-Gesamtscore negativ mit der SUS-Bewertung ($\rho = -0,46, p = ,005$), in der UI-Bedingung zeigte sich ein ähnliches Muster ($\rho = -0,51, p = ,001$).

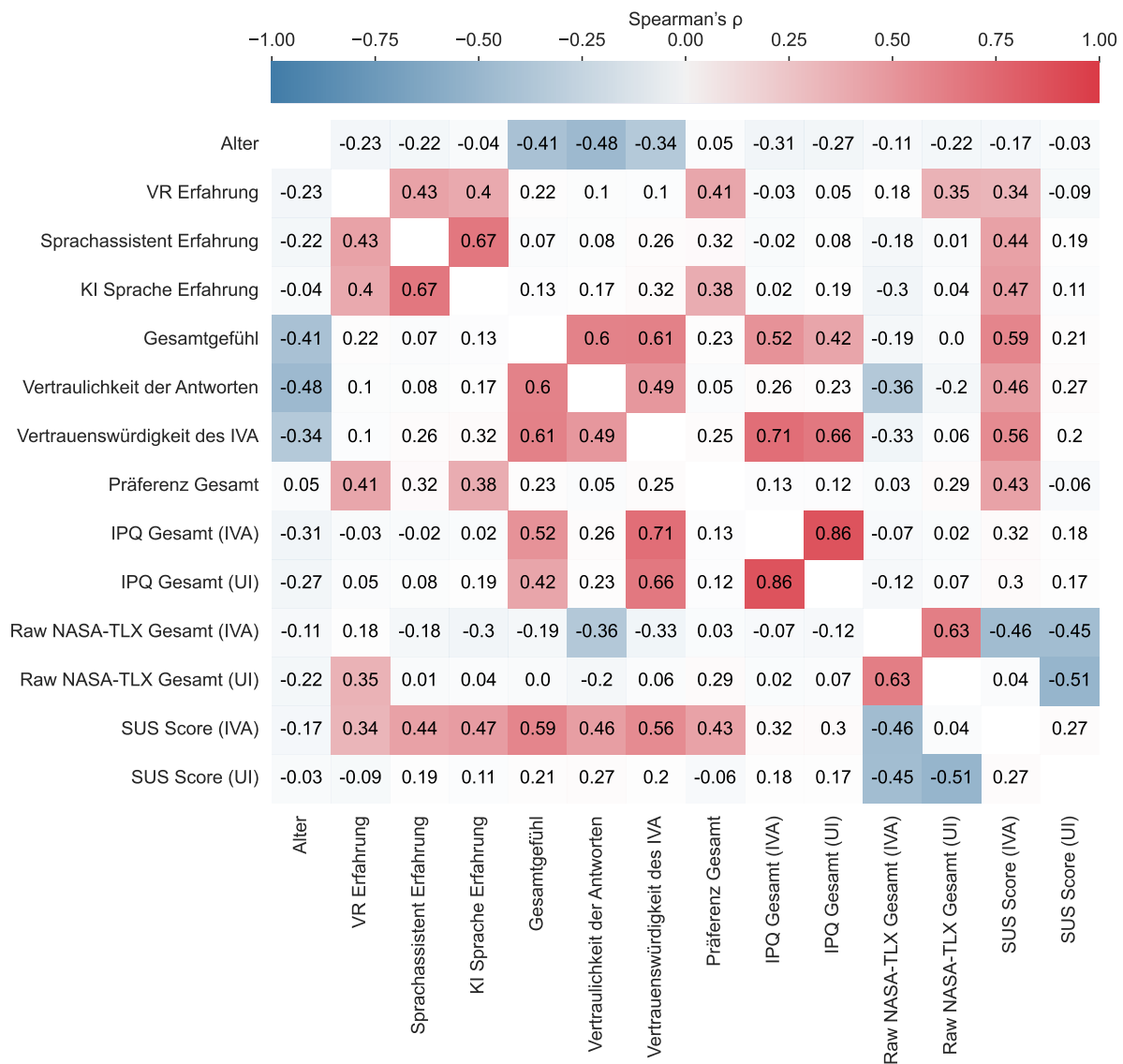


Abbildung 6.14: Korrelationsmatrix zwischen den zentralen Skalen (Spearman- ρ). Signifikante Zusammenhänge ($p < ,05$) sind farblich hervorgehoben, wobei positive Korrelationen in Rot und negative in Blau dargestellt sind.

Erfahrung und Präferenz

Die **Präferenz für die IVA-Interaktion** stand in positivem Zusammenhang mit der Vorerfahrung im Umgang mit VR- und Sprachsystemen. Signifikante Korrelationen ergaben sich zwischen VR-Erfahrung und Präferenz ($\rho = .41, p = ,012$), Sprachassistent-Erfahrung ($\rho = .32, p = ,055$) und KI-Spracherfahrung ($\rho = .38, p = ,022$).

Demografische Einflüsse

Das **Alter** korrelierte negativ mit mehreren Bewertungsaspekten: Ältere Teilnehmende berichteten von einem geringeren Gesamtgefühl während der Befragung ($\rho = -0,41, p = ,012$), einer niedrigeren wahrgenommenen Vertraulichkeit ($\rho = -0,48, p = ,003$) und einem geringeren Vertrauen in den IVA ($\rho = -0,34, p = ,040$). Für Geschlecht und Persönlichkeitseigenschaften (Introversion oder Extraversion) ergaben sich keine signifikanten Korrelationen.

Technische Parameter

Es ergaben sich keine signifikanten Zusammenhänge zwischen technischen Leistungsparametern (z. B. mittlere Latenz, Fehlerhäufigkeit) und den erhobenen psychometrischen Skalen (Präsenz, Arbeitsbelastung, Gebrauchstauglichkeit, Präferenz). Ebenso zeigten sich keine Korrelationen zwischen technischen Messgrößen und demografischen Variablen.

Zusammenfassung

Die Korrelationsanalyse zeigt monotone Zusammenhänge zwischen mehreren der untersuchten Variablen. Positive Beziehungen bestanden zwischen Präsenz, Vertrauen und Usability, während zwischen wahrgenommener Arbeitsbelastung und Usability negative Zusammenhänge beobachtet wurden. Alter und Vorerfahrung korrelierten mit einzelnen Bewertungsdimensionen, technische Kennwerte zeigten hingegen keine signifikanten Beziehungen zu den subjektiven Einschätzungen.

7 Diskussion

Die vorliegende Studie untersuchte die sprachbasierte Fragebogenerhebung durch einen intelligenten virtuellen Agenten (IVA) in Virtual Reality (VR) und verglich sie mit der klassischen Eingabe über eine 2D-Benutzeroberfläche (UI). Im Fokus standen die technische und subjektive Zuverlässigkeit, die Datenäquivalenz der Antworten, die Nutzererfahrung sowie das Präsenzerleben. Ziel war es, zu prüfen, ob die sprachbasierte Erhebung in VR eine valide und praktikable Alternative zu herkömmlichen in-VR-Fragebögen darstellt.

In diesem Kapitel werden die in Kapitel 6 präsentierten empirischen Ergebnisse entlang der Forschungsfragen und Hypothesen zunächst zusammengefasst und anschließend im Kontext des aktuellen Forschungsstandes diskutiert und eingeordnet.

7.1 Zusammenfassung der Ergebnisse

7.1.1 Zuverlässigkeit

Wie zuverlässig funktioniert die Beantwortung von Fragebögen durch einen IVA in VR aus technischer und subjektiver Perspektive?

Technische Zuverlässigkeit

Das System zeigte eine insgesamt stabile Leistung. Über 94 % der 1159 Gesprächsrunden verliefen fehlerfrei und kein Datensatz musste ausgeschlossen werden. Nur acht Ereignisse erforderten einen manuellen Eingriff oder Neustart. Die ASR erkannte 94,2 % der Äußerungen korrekt; Fehltranskriptionen traten in 5,5 % der Fälle auf, meist durch phonetisch ähnliche Wörter. Etwa ein Drittel der fehlerhaften Transkriptionen konnte das LLM automatisch korrigieren. Die mittlere Gesamtlatenz der Sprachpipeline betrug 3,1 Sekunden, wobei die LLM-Verarbeitung mit 69,4 % den größten Anteil ausmachte. Damit lag das System deutlich über der Literaturgrenze für flüssige Dialoge (< 2 s), blieb aber im praktikablen Bereich (< 4 s).

Subjektive Zuverlässigkeit

Die Teilnehmenden bewerteten die Answerfassung überwiegend positiv: 81 % hielten die Erkennung bei Likert-Skalen für gut oder sehr gut. Auch die Zusammenfassungen der Freitextantworten wurden meist als zutreffend beschrieben. Die Vertrauenswürdigkeit und Vertraulichkeit bewerteten jeweils etwa 45 % positiv, rund ein Drittel neutral. Offene Kommentare spiegelten ein insgesamt positives, aber auch durch gelegentliche Verzögerungen und unnatürlicher Wirkung des Agenten beeinträchtigtes Bild.

7.1.2 Datenäquivalenz und Antwortqualität

Unterscheiden sich die gegebenen Antworten auf Fragebögen, wenn diese in VR durch einen IVA oder über eine klassische 2D-UI ausgefüllt werden?

Quantitative Antworten

Die quantitativen Antwortwerte des Igroup Presence Questionnaire (IPQ) zeigten zwischen den beiden Erhebungsbedingungen keine systematischen Unterschiede. Die Streuungen der Subskalen lagen in beiden Bedingungen auf vergleichbarem Niveau ohne signifikante Varianzunterschiede. Auch hinsichtlich der Verteilungsform zeigten sich keine signifikanten Unterschiede. Nur in der UI-Bedingung traten bei den Subskalen *allgemeines Präsenzemfinden (G1)* und *Räumliche Präsenz (SP)* signifikante Links-Schiefen auf.

Qualitative Antworten

Die gegebenen Freitextantworten in der IVA-Bedingung waren signifikant länger und inhaltlich reichhaltiger ($M_{IVA} = 64,4$, $SD_{IVA} = 45,8$ Wörter; $M_{UI} = 18,5$, $SD_{UI} = 16,4$ Wörter; $p < ,001$, $r = -,87$). Ebenso wurden in der IVA-Bedingung mehr kodierte Themen identifiziert ($M_{IVA} = 6,36$, $SD_{IVA} = 2,15$; $M_{UI} = 4,50$, $SD_{UI} = 1,66$, $p < ,001$, $r = -,77$). Auch der Anteil inhaltlich gehaltvoller Antworten war mit 60,2 % höher als bei der UI (47,2 %). Die inhaltliche Analyse zeigte zudem, dass IVA-Antworten häufiger positive und reflektierte Aussagen enthielten, während UI-Antworten meist funktional oder knapp blieben.

7.1.3 Nutzererlebnis

Wie wird die Nutzererfahrung (UX) bei der Befragung durch einen IVA im Vergleich zur Interaktion über eine klassischen 2D-UI wahrgenommen?

Arbeitsbelastung bei Likert-Skalen-Fragen

Die subjektive Arbeitsbelastung (Raw NASA-TLX) zeigte beim Beantworten von Likert-Skalen, dass Teilnehmende in der IVA-Bedingung eine höhere Frustration empfanden ($M_{IVA} = 32,1$, $SD_{IVA} = 28,1$; $M_{UI} = 16,2$, $SD_{UI} = 17,9$; $p < ,001$, $r = -,70$). Die übrigen Dimensionen (Geistige, Körperliche und Zeitliche Anforderung, Leistung und Anstrengung) sowie der Gesamtmittelwert ($M_{IVA} = 26,4$, $SD_{IVA} = 19,8$; $M_{UI} = 19,1$, $SD_{UI} = 15,6$; $p = ,011$, $r = -,45$) zeigten nach Holm-Bonferroni-Korrektur keine signifikanten Unterschiede.

Arbeitsbelastung bei Freitextfragen

Bei Freitextfragen war die subjektive Arbeitsbelastung (Raw NASA-TLX) in der UI-Bedingung signifikant höher ($M_{UI} = 37,6$, $SD_{UI} = 23,8$; $p = ,002$, $r = -,55$), insbesondere in den Dimensionen Körperliche Anforderung ($p < ,001$, $r = -,79$), Leistung ($p = ,002$, $r = -,55$) und Frustration ($p = ,004$, $r = -,54$). Die Sprachinteraktion wurde somit für Freitextfragen als weniger anstrengend und angenehmer erlebt.

Gebrauchstauglichkeit

Die System-Usability-Scale (SUS) zeigte höhere Werte für die IVA-Bedingung als für die UI ($M_{IVA} = 79,3$, $SD_{IVA} = 16,8$; $M_{UI} = 71,0$, $SD_{UI} = 16,4$; $p = ,006$, $r = -,49$), beide über dem Normwert von 70. Nach Bangor et al. (2008) liegen beide Bewertungen im „akzeptablen“ Bereich, wobei die IVA-Bedingung bereits als „gut“ einzustufen ist.

7.1.4 Präsenzerleben

Wie wirkt sich die Befragung durch einen IVA im Vergleich zur Interaktion über eine 2D-UI auf das subjektive Präsenzerleben in der VR-Umgebung aus?

Präsenzerleben

Das subjektive Präsenzerleben, gemessen mit dem IPQ, unterschied sich nicht signifikant zwischen den Erhebungsarten (IPQ-Gesamtscore: $M_{IVA} = 3,36$, $SD_{IVA} = 0,85$, $M_{UI} = 3,32$, $SD_{UI} = 0,80$, $p = ,715$). Auch die Subskalen zeigten keine signifikanten Unterschiede (alle $p \geq ,134$; alle $|d| \leq 0,26$). Das Präsenzerleben blieb somit weitgehend unabhängig von der Befragungsart.

7.2 Interpretation der Ergebnisse

Die Befunde sprechen insgesamt dafür, dass ein sprachgesteuerter, LLM-gestützter Interview-Agent Fragebogenerhebungen in VR zuverlässig durchführen kann, ohne die Ergebnismaße zu verzerren, und zugleich die Erhebung qualitativer Daten deutlich stärkt. Im Folgenden werden die Resultate entlang der Hypothesen H1–H4 interpretiert und im Kontext bisheriger Forschung eingeordnet.

7.2.1 H1 – Technische und subjektive Zuverlässigkeit

H1a – Technische Genauigkeit *Das System erfasst die quantitativen Antworten der Teilnehmenden ohne systematische Abweichungen.*

Die Sprachpipeline lief in der Praxis robust und mit hoher Erkennungsrate. Das LLM führte die Teilnehmenden weitgehend fehlerfrei durch den Fragebogen und extrahierte sowie übermittelte die numerischen und freien Antworten inhaltlich korrekt an das System. Allerdings war das Modell in der vorliegenden Implementierung nicht in der Lage, den Fragebogen vollständig autonom zu erheben, sondern erforderte eine kontinuierliche Überwachung und Korrektur durch das Kontrollsystem. Schwierigkeiten traten vor allem bei der korrekten Zuordnung von Antworten zu den jeweiligen Items auf: Im Mittel kam es in nahezu jeder Fragebogenerhebung zu einer fehlerhaften Zuordnung, und in rund einem Drittel der Erhebungen wurde ein Item übersprungen. Diese Fehler wurden jedoch vom System zuverlässig erkannt und automatisch korrigiert.

In vier Fällen beendete das LLM die Befragung vorzeitig, obwohl die Systeminstruktionen und die Prompts eindeutige Vorgaben enthielten, übersprungene Items zu wiederholen. Um dieses Problem zu beheben, sollten künftige Arbeiten gezielt verschiedene Varianten der Systeminstruktionen und Promptformate erproben, um die Einhaltung der Befragungslogik zu verbessern.

Das ASR-Modul wies insgesamt eine gute Erkennungsleistung auf, erzeugte jedoch vereinzelt fehlerhafte oder unverständliche Transkriptionen, insbesondere bei kurzen Aussagen wie Zahlen. In einem Fall wurde sogar eine andere gültige Zahl ausgegeben. Die betroffene Person bemerkte den Fehler, unternahm jedoch keine Korrekturversuche. Aufgrund der geringen numerischen Abweichung und der hohen Zahl korrekt transkribierter Antworten hatte dieser Einzelfehler keinen signifikanten Einfluss auf die Ergebnisse. Ein wiederholtes Auftreten solcher Fehltranskriptionen oder eine kleinere Stichprobe könnten die Ergebnisse jedoch verfälschen.

Zur Verbesserung der Erkennung numerischer Antworten in zukünftigen Arbeiten sollte ein ASR-Modell mit stärkerer Fokussierung auf Zahlentranskriptionen eingesetzt werden, etwa durch die Einbeziehung eines ergänzenden Datensatzes mit Zahlenaussagen beim

Modelltraining. Alternativ könnten zwei spezialisierte Modelle verwendet werden: eines für kurze Zahlenaussagen und eines für längere Antworten.

Insgesamt zeigte der Prototyp keine systematischen Abweichungen zwischen den erfassten und den angegebenen Antworten der Teilnehmenden. Damit kann die Hypothese H1a bestätigt werden.

H1b – Subjektive Zuverlässigkeit

Die Teilnehmenden bewerten die Fähigkeit des Agenten, ihre Antworten korrekt zu verstehen, als hoch und empfinden ihn als zuverlässigen Interaktionspartner.

Subjektiv fühlten sich die Teilnehmenden überwiegend korrekt verstanden. Die gelegentliche Zurückhaltung bei den Dimensionen Vertrauen und Vertraulichkeit entspricht den Befunden von Ali et al. (2025) und Bäckström (2025), wonach Sprachverarbeitung und KI-basierte Systeme das Bewusstsein hinsichtlich Datenschutz und Datenverwendung erhöhen können. Mehrere Teilnehmende äußerten Unsicherheit und Unwissen darüber, wie ihre Antworten, insbesondere durch die LLM-Komponente, gespeichert oder weiterverarbeitet werden. Für künftige Studien wird daher empfohlen, auf Drittanbieter-Dienste möglichst zu verzichten und stattdessen lokale LLM- und TTS-Modelle einzusetzen. Zudem sollte im Vorgespräch oder zu Beginn des Gesprächs mit dem IVA transparent über sämtliche Aspekte der Datenverarbeitung und des Datenschutzes informiert werden.

Einige Teilnehmende berichteten, sich durch die menschliche Darstellung des Agenten getäuscht zu fühlen, da diese den Eindruck erwecke, es handele sich nicht um eine Maschine. Mehrfach wurde daher der Wunsch nach einer robotischeren Darstellung geäußert, um Missverständnisse zu vermeiden. Zudem empfanden mehrere Personen den Agenten als unecht oder unsympathisch, was den Aufbau von Vertrauen und Gesprächsbereitschaft beeinträchtigte. Für zukünftige Arbeiten wird empfohlen, systematisch zu untersuchen, wie sich die visuelle Gestaltung der Umgebung, des Agenten und seines Verhaltens auf Vertrauen, Rapport und letztlich auf die Datenqualität auswirken. Sofern der zeitliche Rahmen dies zulässt, könnte ein längerer Gesprächseinstieg vor Beginn der eigentlichen Befragung möglicherweise auch den Aufbau von Rapport und Vertrauen zwischen Teilnehmenden und IVA fördern.

Die Fähigkeit des Agenten, Antworten auf Likert-Fragen korrekt zu erfassen, wurde insgesamt hoch eingeschätzt. Auch das Formulieren von Freitextantworten fiel den meisten Teilnehmenden leicht, und die vom IVA generierten Zusammenfassungen wurden als zutreffend wahrgenommen. Diese Ergebnisse stützen die Annahme der Hypothese H1b, dass der IVA als zuverlässiger Interaktionspartner wahrgenommen wird.

Die mittlere Gesamtlatenz der Sprachpipeline betrug rund 3,1 Sekunden. Dieser Wert liegt deutlich oberhalb der in der Literatur häufig genannten Schwelle von etwa zwei Sekunden, innerhalb derer Gesprächspausen noch als unstörend wahrgenommen werden (Miller, 1968; Starner, 2001). Gemäß der Studie von Starner (2001) können längere Verzögerungen zu einer sinkenden Interaktionsbereitschaft und einer erhöhten Frustration führen. Letzteres konnte auch in dieser Studie beobachtet werden.

Gleichzeitig wurde in den qualitativen Kommentaren von den Teilnehmenden angemerkt, dass die visuellen Füller-Animationen des Agenten das subjektive Gefühl der langen Antwortzeit abmilderten. Diese Ergebnisse stehen im Einklang mit den Befunden von Maslych, Katebi et al. (2025), die zeigten, dass konversationelle oder gestische Füller (z. B. Blickbewegungen, Nicken oder Zwischenbemerkungen) die wahrgenommene Wartezeit reduzieren können.

Die Ergebnisse zeigen, dass das verwendete Modell, obwohl es zum Zeitpunkt der Studie zu den leistungsfähigsten Echtzeitmodellen von OpenAI zählte, ohne Kontrollmechanismen nicht hinreichend zuverlässig agierte. Daher musste der Fortschritt der Fragebogenerhebung nicht wie ursprünglich vorgesehen vom LLM, sondern vom lokalen System gesteuert werden. Darüber hinaus war das LLM für den Großteil der gemessenen Latenz verantwortlich. Für zukünftige Entwicklungen eines IVA-basierten Erhebungssystems empfiehlt es sich daher, die Fragebogenerhebungslogik stärker auf lokale Verarbeitung auszurichten und die Abhängigkeit vom LLM zu reduzieren.

So könnten bei Zahlenantworten zu Likert-Items sowohl die Verarbeitung als auch die Generierung kurzer Rückmeldungen (z. B. „Ich habe ‚Fünf‘ verstanden – richtig?“) regelbasiert erfolgen, was zudem die Latenz in solchen Gesprächsrunden um bis zu 70 % verringern könnte. Auch die Formulierung der Item-Fragen ließe sich regelbasiert, etwa mithilfe vordefinierter Satzbausteine, ohne LLM-Unterstützung umsetzen.

Das LLM käme dann gezielt nur bei komplexen oder semantisch anspruchsvollen Eingaben zum Einsatz. Gleichzeitig ließe sich die subjektiv empfundene Reaktionszeit durch natürliche Füllsätze überbrücken (Maslych, Katebi et al., 2025). In einem solchen hybriden Ansatz müsste das Modell zudem nicht den gesamten Fragebogenkontext, sondern lediglich die jeweils aktuelle Interaktionssituation berücksichtigen, was den Einsatz kleinerer, lokaler Modelle ermöglichen würde. Ein solches Verfahren könnte die technische Zuverlässigkeit erhöhen, Latenzen reduzieren und zugleich die Betriebskosten senken.

7.2.2 H2 – Datenäquivalenz und Antwortqualität

H2a – Quantitative Antworten

Die in der IVA-Bedingung erhobenen quantitativen Antwortwerte weisen eine größere Streuung und eine stärker von der Normalverteilung abweichende Verteilungsform auf als in der UI-Bedingung.

Die Ergebnisse zeigen, dass die in VR erhobenen quantitativen Werte (IPQ) zwischen der IVA- und der UI-Erhebung weitgehend äquivalent ausfielen. Die Streuungen unterschieden sich nicht signifikant, und auch die Verteilungsform blieb zwischen beiden Bedingungen stabil. Lediglich in der UI-Bedingung traten bei zwei Subskalen (*General Presence* und *Spatial Presence*) leichte, aber statistisch signifikante Links-Schiefen auf. Für die Gesamtheit der Werte zeigte sich jedoch kein signifikanter Unterschied der Schiefe zwischen den beiden Erhebungsarten. Damit konnte keine systematische Verzerrung der Antwortverteilungen durch die Sprachinteraktion festgestellt werden. Die Hypothese H2a, wonach die IVA-Bedingung zu größeren Streuungen oder zu stärker von der Normalverteilung abweichenden Verteilungsformen führt, wird somit verworfen.

H2b – Qualitative Antworten

Sprachlich gegebene Freitextantworten in der IVA-Bedingung sind ausführlicher und inhaltlich reichhaltiger als über die UI eingegebene Antworten.

Bei offenen Fragen zeigten sich deutliche Vorteile der sprachbasierten Interaktion gegenüber der Texteingabe über die Benutzeroberfläche. Die Freitextantworten in der IVA-Bedingung waren signifikant länger und thematisch vielfältiger. Damit wird die Hypothese H2b, wonach sprachlich gegebene Freitextantworten in der IVA-Bedingung ausführlicher und inhaltlich reichhaltiger sind als über die UI eingegebene Antworten, bestätigt. Diese Befunde legen nahe, dass die Schwelle zur verbalen Formulierung spontaner Gedanken geringer ist als beim Tippen in VR. Ähnliche Effekte beschreiben Alexandrovsky et al. (2021) und Knierim et al. (2018), die auf die hohe mentale Belastung durch virtuelle Tastaturen hinweisen. Darüber hinaus könnten die intuitive Bedienung und soziale Faktoren, wie etwa Rapport, zu reichhaltigeren und längeren Antworten beigetragen haben, wie Gratch und Lucas (2021) und Soni et al. (2022) nahelegen.

7.2.3 H3 – Nutzererfahrung

H3a – Arbeitsbelastung bei Likert-Skalen

Die subjektive Arbeitsbelastung bei der Beantwortung von Likert-Skalen ist in der IVA-Bedingung höher als in der UI-Bedingung.

Die subjektive Nutzererfahrung zeigte ein differenziertes Bild. Bei den Likert-Skalen wiesen die wahrgenommenen Arbeitsbelastungen lediglich in der Dimension *Frustration* eine signifikant höhere Ausprägung in der IVA-Bedingung auf. Dies könnte darauf zurückzuführen sein, dass die Antwortlatenzen häufig über der „Two-Second-Rule“ lagen und sich Aussagen wiederholten, was bei den Teilnehmenden Frustration ausgelöst haben könnte. Auch die Dimension *Geistige Anstrengung* zeigte in der IVA-Bedingung höhere Werte, die

nach Holm-Bonferroni-Korrektur jedoch nicht als signifikant einzustufen waren. Dies deutet auf einen leicht erhöhten kognitiven Aufwand hin, was mit den Ergebnissen von Vankit et al. (2025) übereinstimmt. Die Autoren beschreiben multimodale Darstellungen (Sprache und Visualisierung) als grundsätzlich verständnisfördernd, zugleich aber auch als kognitiv fordernd.

Der Gesamtwert des Raw-NASA-TLX wies keine signifikanten Unterschiede zwischen den Bedingungen auf. Die Hypothese H3a, wonach die subjektive Arbeitsbelastung bei der Beantwortung von Likert-Skalen in der IVA-Bedingung höher ist als in der UI-Bedingung, konnte somit nicht bestätigt werden.

H3b – Arbeitsbelastung bei Freitextfragen

Die subjektive Arbeitsbelastung bei der Beantwortung von Freitextfragen ist in der IVA-Bedingung niedriger als in der UI-Bedingung.

Bei den Freitextfragen zeigte sich hingegen ein entgegengesetztes Muster: Teilnehmende berichteten in der UI-Bedingung von signifikant höheren Anforderungen in den Subskalen *Körperliche Anstrengung*, *Leistung* und *Frustration*. Besonders die körperliche Anstrengung wies eine hohe Effektstärke auf. Auch der Gesamtwert des Raw-NASA-TLX war signifikant höher. Damit wird die Hypothese H3b, wonach die subjektive Arbeitsbelastung bei der Beantwortung von Freitextfragen in der IVA-Bedingung niedriger ist als in der UI-Bedingung, bestätigt. Insgesamt wurde die Spracheingabe als weniger belastend empfunden als die Texteingabe über eine virtuelle Tastatur. Die Befunde stehen im Einklang mit früheren Arbeiten, die eine hohe physische Belastung beim Tippen in VR nachweisen (Alexandrovsky et al., 2021; Knierim et al., 2018). Die signifikant höheren Frustrationswerte sowie die schlechter bewertete Leistungsdimension in der UI-Bedingung könnten zudem darauf hindeuten, dass Teilnehmende mit ihren über die Tastatur eingegebenen Antworten weniger zufrieden waren.

H3c – Gebrauchstauglichkeit

Die subjektive Gebrauchstauglichkeit der sprachbasierten Befragung in der IVA-Bedingung wird höher bewertet als die Interaktion in der UI-Bedingung.

Hinsichtlich der Gebrauchstauglichkeit erzielte die IVA-Bedingung signifikant höhere SUS-Werte mit großer Effektstärke. Dies lässt darauf schließen, dass die sprachbasierte Interaktion insgesamt als benutzerfreundlicher wahrgenommen wurde. Teilnehmende beschrieben die IVA-Befragung häufig als intuitiv, angenehm und unterhaltsam, kritisierten jedoch die eingeschränkte Möglichkeit zur Korrektur und die teilweise langen Antwortverzögerungen. Dass einige Teilnehmende nicht wussten, dass Korrekturen auch verbal möglich gewesen wären, verdeutlicht, dass eine gezielte Einweisung in den Funktionsumfang des IVA die Benutzerfreundlichkeit und das Gefühl von Kontrolle hätte verbessern können.

Die Präferenzdaten spiegeln diese geteilten Ergebnisse wider: Bei Likert-Skalen wurde die UI wegen der schnelleren und präziseren Eingabe bevorzugt, während über drei Viertel der Teilnehmenden für Freitextfragen den IVA vorzogen. Insgesamt bevorzugten rund 60 % der Teilnehmenden den IVA als Gesamtmethode, allerdings ohne signifikanten Unterschied zur UI.

Damit wird die Hypothese H3c, wonach die subjektive Gebrauchstauglichkeit der sprachbasierten Befragung in der IVA-Bedingung höher ist als in der UI-Bedingung, bestätigt, allerdings mit der Einschränkung, dass die Akzeptanz stark vom Fragetyp abhängt. Für die Praxis bietet sich daher ein hybrider Ansatz an, der Sprache für offene und UI-Elemente für geschlossene Fragen kombiniert.

7.2.4 H4 – Präsenzerleben

Das subjektive Präsenzerleben unterscheidet sich nicht signifikant zwischen der IVA- und der UI-Bedingung.

Das subjektive Präsenzerleben unterschied sich weder im IPQ-Gesamtscore noch in den Subskalen signifikant zwischen den Erhebungsarten. Die Ergebnisse legen somit nahe, dass die Präsentationsform des Fragebogens keinen Einfluss auf das berichtete Präsenzerleben hat. Beide Varianten vermieden einen „Break in Presence“ durch das Absetzen des HMD und hielten die Präsenz auf vergleichbarem Niveau. Diese Befunde decken sich mit den Ergebnissen von Safikhani et al. (2021), wonach Interface-Variationen innerhalb der VR kaum Einfluss auf das Präsenzniveau haben, sowie mit jenen von Butz et al. (2022) und Freiwald et al. (2021), die zeigen, dass unterschiedliche Agentendarstellungen das Präsenzerleben meist nicht beeinflussen. Die Hypothese H4, wonach sich das subjektive Präsenzerleben nicht signifikant zwischen der IVA- und der UI-Bedingung unterscheidet, konnte somit bestätigt werden.

Während der Studie zeigten sich jedoch interessante Unterschiede im Verhalten der Teilnehmenden: In der UI-Bedingung richteten sich die Blicke überwiegend auf die Benutzeroberfläche, während Teilnehmende im Gespräch mit dem IVA häufiger umherschauten und auch die VR-Umgebung hinter sich betrachteten. Besonders beim Item „Ich hatte das Gefühl, dass die virtuelle Umgebung hinter mir weitergeht.“ (SP1) fiel dieses Verhalten auf. Obwohl sich daraus in dieser Studie keine signifikanten Antwortunterschiede ergaben, könnte dieses explorative Muster darauf hindeuten, dass die IVA-Interaktion die Aufmerksamkeit stärker auf die aktuell wahrgenommene Umgebung lenkt, anstatt auf die Erinnerung an die zuvor erlebte VR-Experience. Dieser Aspekt sollte in zukünftigen Arbeiten gezielt und systematisch untersucht werden, um mögliche Effekte, beispielsweise auf das Präsenzerleben, zu erfassen.

7.3 Limitationen und Ausblick

Die Befunde dieser Studie sind vor dem Hintergrund mehrerer methodischer und technischer Einschränkungen zu betrachten.

Stichprobe und Generalisierbarkeit

Die Stichprobe ist in Größe und Zusammensetzung nicht repräsentativ. Eine persönliche Bekanntschaft einzelner Teilnehmender mit der Studienleitung sowie ein möglicher Neuartigkeitsfaktor könnten insbesondere Präferenzangaben beeinflusst haben. Zudem umfasste die Stichprobe keine Personen mit motorischen, sensorischen oder kognitiven Einschränkungen, wodurch mögliche Barrieren oder Unterschiede in der Interaktionserfahrung unberücksichtigt blieben. Die Übertragbarkeit der Ergebnisse auf andere Zielgruppen, Sprachen und Anwendungskontexte ist daher eingeschränkt.

Designbedingte Einflüsse

Trotz Ausbalancierung lassen sich Reihenfolge- und Lerneffekte nicht vollständig ausschließen. Da der IVA bereits während der VR-Erfahrung sichtbar war, könnten Teilnehmende, die zunächst die UI-Bedingung absolvierten, von der Anwesenheit des Agenten in der IVA-Bedingung überrascht oder irritiert gewesen sein. Solche situativen Unterschiede könnten den Vergleich der Bedingungen beeinflusst haben. Ebenso könnten Erwartungen an Funktionsweise oder Bedienung, die im ersten Durchlauf entstanden, den zweiten Durchlauf geprägt haben. Darüber hinaus wurden die Items in der IVA-Bedingung auf einer nicht interaktiven Benutzeroberfläche neben dem Agenten visualisiert, um die kognitive Belastung beim Verstehen der gesprochenen Fragen zu reduzieren. Es kann jedoch nicht ausgeschlossen werden, dass diese visuelle Unterstützung die Ergebnisse beeinflusst hat.

Stimulus und Setting

Die Ergebnisse basieren auf einem kurzen, spezifischen VR-Szenario, einem festgelegten Agenten-Design und einer einzelnen 2D-UI. Untersucht wurden ausschließlich Likert-Skalen und offene Fragen, wodurch andere Fragetypen oder Interaktionsformen (z. B. Multiple Choice, Rangordnungen oder adaptive Dialogstrukturen) unberücksichtigt blieben. Andere Szenarien oder Hardware-Setups (z. B. Standalone-Headsets) könnten zudem zu anderen Resultaten führen.

Messinstrumente

Für die Untersuchung des Präsenzerlebens und der Datenäquivalenz wurde derselbe Fragebogen (IPQ) verwendet. Dies reduzierte den zeitlichen Aufwand, führte jedoch zu einer methodischen Kopplung beider Analysen, da identische Items sowohl zur Bewertung des Erlebens als auch zum Vergleich der Antwortäquivalenz herangezogen wurden. Dadurch könnten inhaltliche Überschneidungen die Unabhängigkeit der Befunde teilweise einschränken. Der

NASA-TLX wurde in der ungewichteten Raw-Variante eingesetzt, wodurch individuelle Gewichtungen der Dimensionen unberücksichtigt blieben. Dies kann die Vergleichbarkeit mit anderen Studien einschränken.

ASR und Latenz

Da keine manuellen Referenztranskripte (*Ground-Truth*) der Sprachaufnahmen vorlagen, konnte keine klassische *Word Error Rate* berechnet werden. Die berichtete Erkennungsrate basiert daher auf Plausibilitätsprüfungen der erfassten Antworten und kann insbesondere bei kurzen Äußerungen abweichen. Zudem lag die mittlere Gesamtlatenz der Sprachpipeline über der Komfortschwelle von etwa zwei Sekunden, was Frustration oder ein verstärktes Belastungsempfinden verursacht haben könnte. Bei Systemen mit geringerer Latenz könnten solche Effekte entsprechend schwächer ausfallen.

Implementierung

Die Ergebnisse beziehen sich auf eine spezifische Pipeline (LLM-Version, Prompting, TTS-Stimme). Änderungen an diesen Komponenten könnten Verständlichkeit, Vertrauen oder Antwortverhalten beeinflussen. Da das lokale Kontrollsystem wiederholt korrigierend eingriff, ist die beobachtete Stabilität nicht ausschließlich dem LLM zuzuschreiben.

Qualitative Auswertung

Die Freitextauswertung erfolgte LLM-gestützt mit anschließender manueller Kontrolle, jedoch ohne formale Überprüfung der Übereinstimmung. Subjektive Einflüsse auf die Themenzuordnung und Interpretation können daher nicht ausgeschlossen werden.

Datenschutz und Darstellung

Die Nutzung externer KI-Dienste kann Datenschutzbedenken auslösen. Um Verzerrungen zu vermeiden, wurden technische Details im Vorgespräch bewusst nicht vollständig erläutert.

7.4 Ausblick

Die Ergebnisse dieser Arbeit zeigen, dass sprachbasierte IVA-Befragungen in Virtual Reality eine technisch und methodisch tragfähige Alternative zu klassischen Eingabeverfahren darstellen können. Insbesondere für offene Fragen bietet die Sprachinteraktion deutliche Vorteile hinsichtlich Antwortumfang und inhaltlicher Vielfalt, ohne die Struktur quantitativer Antwortverteilungen zu verzerren. Zugleich wird deutlich, dass Faktoren wie Antwortlatenz, Agentengestaltung und Vertrauen eine wesentliche Rolle für die Akzeptanz und Qualität der Interaktion spielen.

Für zukünftige Arbeiten ergibt sich daraus ein doppelter Handlungsbedarf: einerseits die Optimierung der technischen Pipeline, andererseits die Vertiefung des Verständnisses sozialer und kognitiver Prozesse während der IVA-Interaktion. Technisch sollte der Fokus auf einer

Reduktion der Latenz durch lokal ausgeführte Sprachverarbeitung liegen, um den Gesprächsfluss zu verbessern und Datenschutzbedenken zu minimieren.

Auf inhaltlicher Ebene wäre es lohnend, weitere Fragetypen und Interaktionsformate einzu beziehen, wie Multiple-Choice-Items oder multimodale Antwortoptionen. Auch verschiedene Agentendarstellungen sollten systematisch miteinander verglichen werden, um ihre Effekte auf Vertrauen, Präsenz und Datenqualität besser zu verstehen.

Darüber hinaus wäre es sinnvoll, die IVA-Interaktion über längere Zeiträume oder in realen Anwendungskontexten zu evaluieren, um die Validität der Befunde zu prüfen. Ergänzend sollten künftig auch Personen mit körperlichen, sensorischen oder kognitiven Einschränkungen in die Stichprobe einbezogen werden, um Barrieren und Inklusionspotenziale sprachbasierter Systeme gezielter zu untersuchen.

Insgesamt verdeutlichen die vorliegenden Ergebnisse, dass die Kombination aus Sprache, KI und VR ein hohes Potenzial für die Weiterentwicklung empirischer Erhebungsverfahren bietet. Sie ermöglicht eine neue Form der Interaktion, in der Befragungen weniger als formaler Prozess, sondern als natürliches Gespräch erlebt werden können. Da bislang keine vergleichbaren Ansätze zur sprachbasierten Fragebogenerhebung in VR vorliegen, bilden die Befunde dieser Arbeit eine erste Grundlage für die Weiterentwicklung intelligenter virtueller Agenten, hin zu natürlicheren und zugänglicheren dialogbasierten Befragungen in immersiven virtuellen Umgebungen.

8 Fazit

In der vorliegenden Arbeit wurde untersucht, inwieweit ein sprachgesteuerter, intelligenter virtueller Interview-Agent (IVA) auf Basis eines Large Language Models (LLMs) eine valide und praktikable Alternative zur klassischen in-VR-Fragebogenerhebung über eine 2D-Benutzeroberfläche (UI) darstellt. Im Fokus der Untersuchung standen vier Forschungsfragen: die technische und subjektive Zuverlässigkeit, die Datenäquivalenz der erhobenen Antworten, das Nutzererlebnis während der Erhebung sowie mögliche Einflüsse auf das Präsenzerleben.

Die Befunde zeigen, dass die sprachbasierte Erhebungsmethode in wesentlichen Punkten mit der klassischen UI-Erhebung vergleichbar ist und keine systematischen Verzerrungen der erhobenen Daten verursacht.

Die quantitativen Antworten des *Igroup Presence Questionnaire* (IPQ) fielen in beiden Erhebungsformen weitgehend äquivalent aus; weder die Mittelwerte noch die Streuungen oder die Verteilungsformen unterschieden sich signifikant. Das Präsenzerleben blieb somit stabil, unabhängig davon, ob die Fragen gesprochen oder über die Benutzeroberfläche beantwortet wurden. Dies legt nahe, dass die Sprachinteraktion mit einem IVA in VR eine valide Methode zur quantitativen Fragebogenerhebung darstellt.

Deutliche Unterschiede zeigten sich dagegen in den qualitativen Antworten. Die sprachbasierten Freitextantworten in der IVA-Bedingung waren signifikant länger und thematisch vielfältiger gefasst als die in der UI-Bedingung erfassten Texteingaben. Dies deutet darauf hin, dass die Sprachinteraktion mit einem IVA die Hürden bei offenen Fragen reduziert und somit zu reichhaltigeren qualitativen Daten führt. Damit bietet die IVA-Befragung eine Lösung für die Einschränkung der bisherigen inVRQ-Ansätze, die sich aus Usability-Gründen oft auf geschlossene Skalen beschränkten.

Hinsichtlich des Nutzererlebnisses ergab sich ein differenziertes Bild. Während die sprachbasierte Erhebung bei Likert-Skalenfragen als signifikant frustrierender empfunden wurde, zeigte der *NASA Task Load Index* (Raw NASA-TLX) bei Freitextfragen deutliche Vorteile in der Arbeitsbelastung, vor allem in Bezug auf körperliche Anstrengung, Leistung und Frustration. Diese Asymmetrie erklärt die beobachtete Präferenzstruktur: Likert-Skalen fühlen sich über UI schneller und kontrollierbarer an, Freitextfragen gehen gesprochen leichter von der Hand. Insgesamt bewerteten die Teilnehmenden die IVA-Interaktion in der *System Usability Scale*

(SUS) signifikant positiver als die UI-Bedienung, was auf eine insgesamt höhere Gebrauchstauglichkeit hinweist.

Die technische Umsetzung des Systems zeigte eine robuste Funktionsweise, wenngleich mit hohen Latenzwerten. Die mittlere Gesamtlatenz der Sprachpipeline betrug etwa 3,1 Sekunden und lag somit deutlich über der Zwei-Sekunden-Regel; davon entfielen rund 70 % auf die LLM-Verarbeitung. Die automatische Spracherkennung (ASR) erreichte eine hohe geschätzte Erkennungsrate von rund 95 %, zeigte jedoch Schwächen bei den, für die Fragebogenerhebung relevanten, kurzen Zahläußerungen. Diese Kenngrößen dürften zur erhöhten Frustration bei geschlossenen Fragen beigetragen haben. Das LLM-basierte Fragebogensystem funktionierte insgesamt zuverlässig, erforderte jedoch eine kontinuierliche Überwachung durch das lokale Kontrollsystem. Dies deutet darauf hin, dass derzeit keine vollständig autonome Erhebung über ein LLM realisierbar ist, die ohne Kontrolllogik auskommt.

In Bezug auf den aktuellen Forschungsstand leisten die Ergebnisse dieser Arbeit zwei wesentliche Beiträge. Zunächst konnte empirisch bestätigt werden, dass ein IVA-basierter Ansatz für standardisierte Fragebogenerhebungen in VR grundsätzlich verlässlich ist und dabei keine systematischen Verzerrungen verursacht. Des Weiteren zeigt die Studie, dass die sprachbasierte Interaktion den Umfang und die Vielfalt offener Antworten signifikant verbessert und somit eine methodische Erweiterung klassischer in-VR-Erhebungen darstellt. Diese Befunde untermauern den praktischen Nutzen sprachgesteuerter Agenten in der empirischen VR-Forschung und eröffnen neue Möglichkeiten zur Integration natürlicher Sprache in standardisierte Erhebungsprozesse.

Die in Kapitel 7.3 dargestellten Limitationen verdeutlichen zugleich, dass die Ergebnisse innerhalb eines spezifischen methodischen und technischen Rahmens zu interpretieren sind. Neben der begrenzten Stichprobe betrifft dies insbesondere das kurze VR-Szenario, die konkrete Pipeline-Konfiguration, die spezifischen Fragebogentypen und Agentendarstellung mit Item-Anzeige. Darüber hinaus wurden das Präsenzerleben und die Datenäquivalenz mit demselben Instrument (IPQ) erfasst, was zwar den Erhebungsaufwand reduzierte, aber methodische Abhängigkeiten erzeugt haben kann. Diese Einschränkungen relativieren die Generalisierbarkeit der Befunde, ohne deren grundsätzliche Aussagekraft infrage zu stellen.

Aus den Ergebnissen lassen sich mehrere Empfehlungen für zukünftige Arbeiten ableiten. Ein hybrides Erhebungskonzept, das sprachliche und visuelle Eingabeformen kombiniert, könnte die Stärken beider Ansätze vereinen, würde jedoch die Barrierefreiheit des sprachbasierten Systems einschränken, da dafür wieder zusätzliche Eingabegeräte wie Controller erforderlich wären.

Für zukünftige Implementierungen eines IVA-Fragebogensystems sind zwei grundlegende Anpassungen zu empfehlen: Zum einen die Implementierung schneller, regelbasierter Reaktionen durch deterministische Zahlenerkennung bei Likert-Skalen sowie natürlicher Fülläußerungen und Rückfragen zur Überbrückung von Wartezeiten. Des Weiteren empfiehlt

es sich, die Steuerungslogik stärker lokal auszuführen und alle Modelle lokal zu betreiben, um Abhängigkeiten von Cloud-basierten LLMs zu reduzieren und sowohl Datenschutz als auch Reaktionszeiten zu verbessern. Beide Maßnahmen könnten Antwortzeiten verkürzen, Kosten senken und das Verhalten des Systems verlässlicher und konsistenter machen. So ließen sich die meisten Nachteile der vorliegenden Umsetzung sowie auch die vergleichsweise lange Erhebungsdauer verbessern.

Darüber hinaus bietet es sich an, verschiedene Agentengestaltungen, Stimmlagen und Abstraktionsgrade systematisch zu untersuchen, um deren Einfluss auf quantitative und qualitative Antworten gezielt zu erfassen. Schließlich sollte die Generalisierbarkeit der Ergebnisse in zukünftigen Studien durch größere und diversere Stichproben sowie komplexere VR-Szenarien und weitere Fragebogentypen geprüft werden.

Die vorliegende Studie kommt zu dem Schluss, dass sprachbasierte IVA-Befragungen eine methodisch tragfähige und technisch realisierbare Alternative zu klassischen in-VR-Erhebungsverfahren darstellen. Sie ermöglichen eine zuverlässige Erfassung quantitativer Daten und fördern zugleich die Qualität qualitativer Antworten. Trotz gewisser methodischer und technischer Einschränkungen scheint der Einsatz von Sprachinteraktion eine praktikable Ergänzung etablierter Verfahren der Datenerhebung in immersiven virtuellen Umgebungen zu sein. Perspektivisch können weiterentwickelte IVA-Systeme dazu beitragen, Datenerhebungen in VR effizienter, zugänglicher und nutzerorientierter zu gestalten.

Literatur

- Affara, M., Lagu, H. I., Achol, E., Karamagi, R., Omari, N., Ochido, G., Kezakarayagwa, E., Kabatesi, F., Nkeshimana, A., Roba, A., Ndia, M. N., Abudo, M. U., Kabanda, A., Mpabuka, E., Mwikarago, E. I., Kutjok, P. E., Samson, D. D., Deng, L. L., Moremi, N., ... Gehre, F. (2021). The East African Community (EAC) mobile laboratory networks in Kenya, Burundi, Tanzania, Rwanda, Uganda, and South Sudan—from project implementation to outbreak response against Dengue, Ebola, COVID-19, and epidemic-prone diseases. *BMC Medicine*, 19(1). <https://doi.org/10.1186/s12916-021-02028-y>
- Alexandrovsky, D., Putze, S., Bonfert, M., Höffner, S., Michelmann, P., Wenig, D., Malaka, R., & Smeddinck, J. D. (2020). Examining Design Choices of Questionnaires in VR User Studies. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3313831.3376260>
- Alexandrovsky, D., Putze, S., Schülke, A., & Malaka, R. (2021, 24. April). Towards Low-burden Responses to Open Questions in VR. <https://doi.org/10.48550/arXiv.2104.12020>
- Alghamdi, N., & Cristea, A. I. (2024). Integrating Speech Input in Educational Immersive Virtual Reality Applications: A Systematic Review. *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, 1–8. <https://doi.org/10.1109/IS61756.2024.10705165>
- Ali, M., Arunasalam, A., & Farrukh, H. (2025, 9. April). Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms. <https://doi.org/10.48550/arXiv.2504.06552>
- Andreev, A., & Cattan, G. (2023, 20. Juli). Towards an architectural framework for intelligent virtual agents using probabilistic programming. <https://doi.org/10.48550/arXiv.2307.10693>
- Augustin, A., & Contributors. (2025, 5. März). *Websockets: A Python Library for WebSocket Servers and Clients* (Version 15.0.1). <https://websockets.readthedocs.io/>
- Bäckström, T. (2025, 21. August). Privacy in speech technology. <https://doi.org/10.48550/arXiv.2305.05227>
- Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction*, 24(6), 574–594. <https://doi.org/10.1080/10447310802205776>
- Bareišytė, L., Slatman, S., Austin, J., Rosema, M., Van Sintemaartensdijk, I., Watson, S., & Bode, C. (2024). Questionnaires for evaluating virtual reality: A systematic scoping

- review. *Computers in Human Behavior Reports*, 16, 100505. <https://doi.org/10.1016/j.chbr.2024.100505>
- Berkman, M. I., & Akan, E. (2024). Presence and immersion in virtual reality. Springer, Cham. https://doi.org/10.1007/978-3-031-23161-2_162
- Bickmore, T. W., Pfeifer, L. M., & Paasche-Orlow, M. K. (2009). Using computer agents to explain medical documents to patients with low health literacy. *Patient Education and Counseling*, 75(3), 315–320. <https://doi.org/10.1016/j.pec.2009.02.007>
- Blanz, M. (2021). *Forschungsmethoden und Statistik für die Soziale Arbeit*. W. Kohlhammer GmbH. <https://doi.org/10.17433/978-3-17-039819-1>
- Botella, C., Alcañiz Raya, M., Liaño, V., Guerrero, B., & Rey, B. (2005). Immersion and Emotion: Their Impact on the Sense of Presence. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, 7, 734–41. <https://doi.org/10.1089/cpb.2004.7.734>
- Brooke, J. (1996). SUS - a quick and dirty usability scale. *Usability Evaluation in Industry*.
- Bulmer, M. (1979). *Principles of Statistics*. Dover Publications.
- Butz, M., Hepperle, D., & Wölfel, M. (2022). Influence of visual appearance of agents on presence, attractiveness, and agency in virtual reality. In M. Wölfel, J. Bernhardt & S. Thiel (Hrsg.), *ArtsIT, interactivity and game creation* (S. 44–60). Springer International Publishing. https://doi.org/10.1007/978-3-030-95531-1_4
- Castillo-López, G., de Chalendar, G., & Semmar, N. (2025, Mai). A Survey of Recent Advances on Turn-taking Modeling in Spoken Dialogue Systems. In M. I. Torres, Y. Matsuda, Z. Callejas, A. del Pozo & L. F. D’Haro (Hrsg.), *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology* (S. 254–271). Association for Computational Linguistics. <https://aclanthology.org/2025.iwds-1.27/>
- Chang, Z., Bai, H., Zhang, L., Gupta, K., He, W., & Billingham, M. (2022). The impact of virtual agents’ multimodal communication on brain activity and cognitive load in Virtual Reality. *Frontiers in Virtual Reality*, 3. <https://doi.org/10.3389/frvir.2022.995090>
- Chheang, V., Sharmin, S., Marquez-Hernandez, R., Patel, M., Rajasekaran, D., Caulfield, G., Kiafar, B., Li, J., Kullu, P., & Barmaki, R. L. (2024, 24. Januar). Towards Anatomy Education with Generative AI-based Virtual Assistants in Immersive Virtual Reality Environments. <https://doi.org/10.48550/arXiv.2306.17278>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Psychology Press. <https://doi.org/10.4324/9780203771587>
- Colabianchi, S., Tedeschi, A., & Costantino, F. (2023). Human-technology integration with industrial conversational agents: A conceptual architecture and a taxonomy for manufacturing. *Journal of Industrial Information Integration*, 35, 100510. <https://doi.org/10.1016/j.jii.2023.100510>
- Craig, A. B., Sherman, W. R., & Will, J. D. (2009, 2. Juni). *Developing virtual reality applications: Foundations of effective design*. Morgan Kaufmann.

- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Cui, W., Yu, D., Jiao, X., Meng, Z., Zhang, G., Wang, Q., Guo, Y., & King, I. (2025, 6. Februar). Recent advances in speech language models: A survey. <https://doi.org/10.48550/arXiv.2410.03751>
- DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hart-holt, A., Lhommet, M., Lucas, G., Marsella, S., Morbini, F., Nazarian, A., Scherer, S., Stratou, G., Suri, A., Traum, D., Wood, R., ... Morency, L.-P. (2014). SimSensei kiosk: a virtual human interviewer for healthcare decision support. *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, 1061–1068. <https://dl.acm.org/doi/10.5555/2615731.2617415>
- Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00026>
- Dreyer, E. (2025). *NativeWebSocket: WebSocket client for Unity*. Verfügbar 2. September 2025 unter <https://github.com/endel/NativeWebSocket>
- Eckstein, P. P., Götze, W., Hartl, F., Rönz, B., & Strohe, H. G. (1994). *Lexikon Statistik* (B. Rönz & H. G. Strohe, Hrsg.). Gabler Verlag. <https://doi.org/10.1007/978-3-322-91144-5>
- Elfleet, M., & Chollet, M. (2024). Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, 1–9. <https://doi.org/10.1145/3652988.3673965>
- Ermí, L., & Mäyrä, F. (2005). Fundamental components of the gameplay experience: Analysing immersion. *Proceedings of DiGRA 2005 Conference: Changing Views: Worlds in Play*. <https://doi.org/10.26503/dl.v2005i1.119>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Feick, M., Kleer, N., Tang, A., & Krüger, A. (2020). The Virtual Reality Questionnaire Toolkit. *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 68–69. <https://doi.org/10.1145/3379350.3416188>
- Ferstl, Y., Thomas, S., Guiard, C., Ennis, C., & McDonnell, R. (2021). Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 76–83. <https://doi.org/10.1145/3472306.3478338>
- Foundation, P. S. (2025). Python.org. <https://www.python.org/>
- Freiwald, J. P., Schenke, J., Lehmann-Willenbrock, N., & Steinicke, F. (2021). Effects of Avatar Appearance and Locomotion on Co-Presence in Virtual Reality Collaborations. *Proceedings of Mensch und Computer 2021*, 393–401. <https://doi.org/10.1145/3473856.3473870>

- Fribourg, R., Argelaguet, F., Lécuyer, A., & Hoyet, L. (2020). Avatar and Sense of Embodiment: Studying the Relative Preference Between Appearance, Control and Point of View. *IEEE Transactions on Visualization and Computer Graphics*, 26(5), 2062–2072. <https://doi.org/10.1109/TVCG.2020.2973077>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18. <https://doi.org/10.1037/a0024338>
- Gao, M., Kortum, P., & Oswald, F. L. (2020). Multi-Language Toolkit for the System Usability Scale. *International Journal of Human–Computer Interaction*, 36(20), 1883–1901. <https://doi.org/10.1080/10447318.2020.1801173>
- Geng, S., Cooper, H., Moskal, M., Jenkins, S., Berman, J., Ranchin, N., West, R., Horvitz, E., & Nori, H. (2025, 27. Februar). JSONSchemaBench: A Rigorous Benchmark of Structured Outputs for Language Models. <https://doi.org/10.48550/arXiv.2501.10868>
- Godoy, D. V. (2021a). *Full Transformer Architecture* [CC BY 4.0]. https://github.com/dvgodoy/dl-visuals/blob/main/Transformers/full_transformer.png
- Godoy, D. V. (2021b). *Recurrent Neural Network (RNN) schematic* [CC BY 4.0]. <https://github.com/dvgodoy/dl-visuals/blob/main/RNNs/rnn2.png>
- Gratch, J., & Lucas, G. (2021, 10. September). Rapport between humans and socially interactive agents. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 433–462). ACM. <https://doi.org/10.1145/3477322.3477335>
- Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine learning*, 369–376. <https://doi.org/10.1145/1143844.1143891>
- Harms, C., & Biocca, F. (2004). Internal consistency and reliability of the networked minds measure of social presence. <https://web-archive.southampton.ac.uk/cogprints.org/7026/>
- Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. <https://doi.org/10.1177/154193120605000909>
- Hart, S. G., & Staveland, L. E. (1988, 1. Januar). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Hrsg.), *Advances in Psychology* (S. 139–183, Bd. 52). North-Holland. Verfügbar 22. August 2025 unter [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hartholt, A., & Mozgai, S. (2022, 27. Oktober). Platforms and tools for SIA research and development. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 261–304). ACM. <https://doi.org/10.1145/3563659.3563668>
- Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., & Gratch, J. (2013). All together now: Introducing the virtual human toolkit. In R. Aylett,

- B. Krenn, C. Pelachaud & H. Shimodaira (Hrsg.), *Intelligent virtual agents* (S. 368–381). Springer. https://doi.org/10.1007/978-3-642-40415-3_33
- hecomi & Contributors. (2025, 21. August). *uLipSync*. Verfügbar 21. August 2025 unter <https://github.com/hecomi/uLipSync>
- Hepperle, D., Weiß, Y., Siess, A., & Wölfel, M. (2019). 2d, 3d or speech? a case study on which user interface is preferable for what kind of object interaction in immersive virtual reality. *Computers & Graphics*, 82, 321–331. <https://doi.org/10.1016/j.cag.2019.06.003>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <http://www.jstor.org/stable/4615733>
- Jaiswal, S., Valstar, M., Kusumam, K., & Greenhalgh, C. (2019). Virtual Human Questionnaire for Analysis of Depression, Anxiety and Personality, 81–87. <https://doi.org/10.1145/3308532.3329469>
- Jansen, B. J., Jung, S.-g., & Salminen, J. (2023). Employing large language models in survey research. *Natural Language Processing Journal*, 4, 100020. <https://doi.org/10.1016/j.nlp.2023.100020>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd). <https://web.stanford.edu/~jurafsky/slp3/>
- Kay, A. (1984). Computer Software. *Scientific American*, 53–59. https://www.academia.edu/113414068/Computer_Software
- Knibbe, J., Schjerlund, J., Petraeus, M., & Hornbæk, K. (2018). The Dream is Collapsing: The Experience of Exiting VR. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3173574.3174057>
- Knierim, P., Schwind, V., Feit, A. M., Nieuwenhuizen, F., & Henze, N. (2018). Physical Keyboards in Virtual Reality: Analysis of Typing Performance and Effects of Avatar Hands. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–9. <https://doi.org/10.1145/3173574.3173919>
- Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024). Careless Whisper: Speech-to-Text Hallucination Harms. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1672–1681. <https://doi.org/10.1145/3630106.3658996>
- Kopp, S., & Hassan, T. (2022, 27. Oktober). The fabric of socially interactive agents: Multimodal interaction architectures. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 77–112). ACM. <https://doi.org/10.1145/3563659.3563664>

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Lacombe, Y. (2024, 10. Januar). *Openai/Whisper-Large-V3-Turbo*. Verfügbar 26. August 2025 unter <https://huggingface.co/openai/whisper-large-v3-turbo>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Leschanowsky, A., Rech, S., Popp, B., & Bäckström, T. (2024). Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review. *Computers in Human Behavior*, 159, 108344. <https://doi.org/10.1016/j.chb.2024.108344>
- Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025, 8. Juni). Exploring the Impact of Temperature on Large Language Models: Hot or Cold? <https://doi.org/10.48550/arXiv.2506.07295>
- LiveKit. (n. d.). *LiveKit agents* [LiveKit docs]. Verfügbar 2. Juli 2025 unter <https://docs.livekit.io/>
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100. <https://doi.org/10.1016/j.chb.2014.04.043>
- Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers. *Frontiers in Robotics and AI*, 4. <https://doi.org/10.3389/frobt.2017.00051>
- Lugrin, B. (2021). Introduction to Socially Interactive Agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition* (1. Aufl., S. 1–20, Bd. 37). Association for Computing Machinery. <https://doi.org/10.1145/3477322.3477324>
- Lugrin, B., Pelachaud, C., & Traum, D. (Hrsg.). (2021, August). *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition* (1. Aufl., Bd. 37). Association for Computing Machinery. <https://doi.org/10.1145/3477322>
- Maslych, M., Katebi, M., Lee, C., Hmaiti, Y., Ghasemaghaei, A., Pumarada, C., Palmer, J., Segarra Martinez, E., Emporio, M., Snipes, W., McMahan, R. P., & LaViola Jr., J. J. (2025). Mitigating response delays in free-form conversations with LLM-powered intelligent virtual agents. *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 1–15. <https://doi.org/10.1145/3719160.3736636>
- Maslych, M., Pumarada, C., Ghasemaghaei, A., & Jr, J. J. L. (2025, 6. Januar). Takeaways from applying LLM capabilities to multiple conversational avatars in a VR pilot study. <https://doi.org/10.48550/arXiv.2501.00168>

- Mathysen, D., & Glorieux, I. (2021). Integrating virtual reality in qualitative research methods: Making a case for the VR-assisted interview. *Methodological Innovations*, 14(2), 20597991211030778. <https://doi.org/10.1177/20597991211030778>
- McDonnell, R., & Mutlu, B. (2021, 10. September). Appearance. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 105–146). ACM. <https://doi.org/10.1145/3477322.3477327>
- Meta Platforms, Inc. (2023). *Meta Quest 3: Mixed-Reality-Headset der nächsten Generation*. Verfügbar 23. Oktober 2025 unter <https://www.meta.com/de/quest/quest-3/>
- Microsoft. (2025a). *Sprach- und Stimmunterstützung für den Speech-Dienst*. Verfügbar 16. Oktober 2025 unter <https://learn.microsoft.com/de-de/azure/ai-services/speech-service/language-support?tabs=tts>
- Microsoft. (2025b, 7. August). *Speech service documentation*. Verfügbar 30. August 2025 unter <https://learn.microsoft.com/en-us/azure/ai-services/speech-service/>
- Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1994). Augmented reality: A class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies*, 2351. <https://doi.org/10.1117/12.197321>
- Miller, R. B. (1968). Response time in man-computer conversational transactions. *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, 267–277. <https://doi.org/10.1145/1476589.1476628>
- Morgan, W. A. (1939). A Test for the Significance of the Difference Between the Two Variances in a Sample From a Normal Bivariate Population. *Biometrika*, 31(1), 13–19. <https://doi.org/10.2307/2334972>
- Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- Mott, M., Tang, J., Kane, S., Cutrell, E., & Ringel Morris, M. (2020). “I just went into it assuming that I wouldn’t be able to have the full experience”: Understanding the Accessibility of Virtual Reality for People with Limited Mobility. *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–13. <https://doi.org/10.1145/3373625.3416998>
- NASA Human Systems Integration Division. (2022). *NASA Task Load Index (NASA-TLX) Paper and Pencil Package*. Verfügbar 1. August 2025 unter <https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf>
- Nielsen, J. (1994, Oktober). *Usability Engineering*. Morgan Kaufmann Publishers Inc.
- OpenAI. (2025a). *What are tokens and how to count them?* [OpenAI help center]. Verfügbar 29. August 2025 unter <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>
- OpenAI. (2025b, 14. April). *Introducing GPT-4.1 in the API*. Verfügbar 19. August 2025 unter <https://openai.com/index/gpt-4-1/>

- OpenAI. (2025c, 11. Oktober). *OpenAI Python API library* (Version 1.90.0). Verfügbar 11. Oktober 2025 unter <https://github.com/openai/openai-python>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, 4. März). Training language models to follow instructions with human feedback. <https://doi.org/10.48550/arXiv.2203.02155>
- Özkaya, S., Berrezueta-Guzman, S., & Wagner, S. (2025, 4. August). How LLMs are Shaping the Future of Virtual Reality. <https://doi.org/10.48550/arXiv.2508.00737>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimselshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: an imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8026–8037. <https://doi.org/10.48550/arXiv.1912.01703>
- Pelachaud, C., Busso, C., & Heylen, D. (2021, 10. September). Multimodal behavior modeling for socially interactive agents. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 259–310). ACM. <https://doi.org/10.1145/3477322.3477331>
- Pietroszek, K. (2018). Raycasting in virtual reality. In *Encyclopedia of computer graphics and games* (S. 1–3). Springer, Cham. https://doi.org/10.1007/978-3-319-08234-9_180-1
- Pitman, E. J. G. (1939). A Note on Normal Correlation. *Biometrika*, 31(1), 9–12. <https://doi.org/10.2307/2334971>
- primeLine AI Services. (2024). Whisper-Large-V3-Turbo-German. <https://doi.org/10.57967/hf/3827>
- Puradiredja, D. I., Boenecke, J., Habermann, T., & eVeeR Projektteam. (2025). eVeeR: Educational virtual environment for epidemic response. Verfügbar 10. Oktober 2025 unter <https://www.bnitm.de/ausbildung/zentrum-fuer-fort-und-weiterbildung/iace/projects/eveer>
- Putze, S., Alexandrovsky, D., Putze, F., Höffner, S., Smeddinck, J. D., & Malaka, R. (2020). Breaking the experience: Effects of questionnaires in VR user studies. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3313831.3376144>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, 6. Dezember). Robust Speech Recognition via Large-Scale Weak Supervision. <https://doi.org/10.48550/arXiv.2212.04356>
- rany2. (2025, 30. August). *Rany2/Edge-TTS*. Verfügbar 30. August 2025 unter <https://github.com/rany2/edge-tts>
- Raux, A. (2008). Flexible turn-taking for spoken dialog systems. *Language Technologies Institute, CMU Dec.*

- Reallusion Inc. (2025a). *iClone: Real-time 3D Character Animation Software* [3D Animation Software for Character Animator | iClone | Reallusion]. Verfügbar 20. August 2025 unter <https://www.reallusion.com/iclone/default.html>
- Reallusion Inc. (2025b). *Realistic Human* [Abbildung]. Verfügbar 20. August 2025 unter <https://www.reallusion.com/de/character-creator/3d-content.html>
- Renze, M., & Guven, E. (2024). The Effect of Sampling Temperature on Problem Solving in Large Language Models, 7346–7356. <https://doi.org/10.18653/v1/2024.findings-emnlp.432>
- Rzeszewski, M., Osborne, T., Jones, P., Evans, L., & Weitkamp, G. (2024). Interviewing in the metaverse: The renewed importance of location and embodiment. *Applied Geography*, 167, 103295. <https://doi.org/10.1016/j.apgeog.2024.103295>
- Safikhani, S., Holly, M., Kainz, A., & Pirker, J. (2021). The influence of in-VR questionnaire design on the user experience. *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, 1–8. <https://doi.org/10.1145/3489849.3489884>
- Safikhani, S., Nacke, L., & Pirker, J. (2024). A Literature Review and Taxonomy of In-VR Questionnaire User Interfaces. <https://doi.org/10.48550/ARXIV.2406.01122>
- Saund, C., & Marsella, S. (2021, 10. September). Gesture generation. In B. Lugrin, C. Pelachaud & D. Traum (Hrsg.), *The handbook on socially interactive agents* (1. Aufl., S. 213–258). ACM. <https://dl.acm.org/doi/10.1145/3477322.3477330>
- Schmidt, S., Rolff, T., Voigt, H., Offe, M., & Steinicke, F. (2024). Natural expression of a machine learning model’s uncertainty through verbal and non-verbal behavior of intelligent virtual agents. *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–15. <https://doi.org/10.1145/3654777.3676454>
- Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The Experience of Presence: Factor Analytic Insights. *Presence: Teleoperators and Virtual Environments*, 10(3), 266–281. <https://doi.org/10.1162/105474601300343603>
- Schuir, J., Pöhler, L., & Teuteberg, F. (2022). Zwischen Preisjägern, Datenschützern und Tech-Enthusiasten: Segmentierung des Virtual-Reality-Marktes am Beispiel Oculus. *HMD Praxis der Wirtschaftsinformatik*, 59(1), 261–279. <https://doi.org/10.1365/s40702-021-00817-w>
- Schwind, V., Knierim, P., Haas, N., & Henze, N. (2019). Using Presence Questionnaires in Virtual Reality. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. <https://doi.org/10.1145/3290605.3300590>
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3), 591–611. <https://doi.org/10.1093/biomet/52.3-4.591>
- Sherman, W., & Craig, A. (2002, 1. Januar). *Understanding virtual reality: Interface, application, and design*.
- Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., & Hagita, N. (2008). How quickly should communication robots respond? *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 153–160. <https://doi.org/10.1145/1349822.1349843>

- Silero Team. (2024). *Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier*. Verfügbar 25. August 2025 unter <https://github.com/snakers4/silero-vad>
- Slater, M. (2003). A Note on Presence Terminology. *Presence Connect*, 3.
- Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *British Journal of Psychology*, 109(3), 431–433. <https://doi.org/10.1111/bjop.12305>
- Slater, M., & Steed, A. (2000). A Virtual Presence Counter. *Presence*, 9, 413–434. <https://doi.org/10.1162/105474600566925>
- Slater, M., & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6(6), 603–616. <https://doi.org/10.1162/pres.1997.6.6.603>
- Soni, H., Ivanova, J., Wilczewski, H., Bailey, A., Ong, T., Narma, A., Bunnell, B. E., & Welch, B. M. (2022). Virtual conversational agents versus online forms: Patient experience and preferences for health data collection. *Frontiers in Digital Health*, 4. <https://doi.org/10.3389/fdgth.2022.954069>
- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Sriram, V. (2025). Understanding the architecture of voice assistants: A technical deep dive. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(2), 587–595. <https://doi.org/10.32628/CSEIT25112398>
- Starner, T. (2001). The Challenges of Wearable Computing: Part 2. *IEEE Micro*, 21(4), 54–67. <https://doi.org/10.1109/40.946683>
- Swerts, M., & Kraemer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1), 81–94. <https://doi.org/10.1016/j.jml.2005.02.003>
- Technologies, U. (2025). Unity Real-Time Development Platform. <https://unity.com/>
- Tepe, T., Kaleci, D., & Tüzün, H. (2024). Virtual reality applications in education. In *Encyclopedia of computer graphics and games* (S. 2007–2013). Springer, Cham. https://doi.org/10.1007/978-3-031-23161-2_166
- Todericiu, I. A. (2025). Virtual assistants: A review of the next frontier in AI interaction. *Acta Universitatis Sapientiae, Informatica*, 17(1), 1. <https://doi.org/10.1007/s44427-025-00002-7>
- Unity Technologies. (2025). *Spatial Keyboard*. Verfügbar 1. September 2025 unter <https://docs.unity3d.com/Packages/com.unity.xr.interaction.toolkit@3.2/manual/samples-spatial-keyboard.html>
- Vankit, S. A., Zamanifard, S., Diaz, D., Mousas, C., Richardson, K., Duchowski, A. T., & Volonte, M. (2025). Exploring the impact of multimodal long conversations in VR on attitudes toward behavior change, memory retention, and cognitive load. *Computer Animation and Virtual Worlds*, 36(3), e70023. <https://doi.org/10.1002/cav.70023>

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, 12. Juni). Attention Is All You Need. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, Z., Rao, M., Ye, S., Song, W., & Lu, F. (2025, 13. Juni). Towards spatial computing: recent advances in multimodal natural interaction for XR headsets. <https://doi.org/10.48550/arXiv.2502.07598>
- Weidner, F., Boettcher, G., Arboleda, S. A., Diao, C., Sinani, L., Kunert, C., Gerhardt, C., Broll, W., & Raake, A. (2023). A Systematic Review on the Visualization of Avatars and Agents in AR & VR displayed using Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics*, 29(5), 2596–2606. <https://doi.org/10.1109/TVCG.2023.3247072>
- Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), 80. <https://doi.org/10.2307/3001968>
- Witmer, B. G., & Singer, M. J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225–240. <https://doi.org/10.1162/105474698565686>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- Xu, Y., Feng, A. W., Marsella, S., & Shapiro, A. (2013). A Practical and Configurable Lip Sync Method for Games. *Proceedings of Motion on Games*, 131–140. <https://doi.org/10.1145/2522628.2522904>
- Yeo, Y. H., Clark, A., Mehra, M., Danovitch, I., Osilla, K., Yang, J. D., Kuo, A., Kim, H.-S., Vipani, A., Wang, Y., Ayoub, W., Trivedi, H., Samaan, J. S., Wu, T., Shah, V. H., Liran, O., & Spiegel, B. (2024). The Feasibility and Usability of an Artificial Intelligence-Enabled Conversational Agent in Virtual Reality for Patients with Alcohol-Associated Cirrhosis: A Multi-Methods Study. *Journal of Medical Extended Reality*, 1(1), 257–270. <https://doi.org/10.1089/jmxr.2024.0033>
- Young, J., Ferreira, J., & Pantidi, N. (2024). „I Shot the Interviewer!“. The Effects of In-VR Interviews on Participant Feedback and Rapport. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17. <https://doi.org/10.1145/3613904.3642707>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2025, 11. März). A survey of large language models. <https://doi.org/10.48550/arXiv.2303.18223>

Benutzte Hilfsmittel

Zur Anfertigung der vorliegenden Masterarbeit wurden neben den im Text und im Literaturverzeichnis genannten Quellen folgende Hilfsmittel verwendet:

- **JATOS (Just Another Tool for Online Studies):** Lokales Hosting und Verwaltung der Fragebögen außerhalb der VR (NASA-TLX, SUS, Abschluss- und demografischer Fragebogen).
- **Overleaf:** Erstellung und Formatierung der Arbeit in \LaTeX .
- **Zotero:** Literaturverwaltung und Quellenorganisation.
- **Python (pandas, scipy, matplotlib):** Statistische Berechnungen und grafische Auswertungen.
- **PyCharm & Rider (JetBrains):** Entwicklung, Debugging und Verwaltung des Quellcodes für Backend- und Unity-Komponenten sowie Ausführung der Backend-Module des Prototyps.
- **Unity (Unity Technologies):** Implementierung und Ausführung der IVA-Integration und Benutzerinteraktion des VR-Prototyps.
- **ChatGPT (OpenAI, Modell GPT-5, 2025):** Unterstützung bei der sprachlichen Überarbeitung von Texten, bei der Erstellung von Tabellen- und \LaTeX -Layouts, bei der Strukturierung von Inhalten, bei der Generierung einzelner Illustrationen sowie bei der Entwicklung und Fehlersuche im Softwareprototyp.
- **LanguageTool:** Überprüfung der Rechtschreibung und Grammatik.
- **DeepL Write / DeepL Translator:** Sprachliche und stilistische Überprüfung einzelner englischer Begriffe und Satzformulierungen.
- **Affinity Designer:** Erstellung und Gestaltung von Abbildungen und Grafiken.
- **Blender:** Anpassung des Agenten-3D-Modells.
- **draw.io (diagrams.net):** Erstellung von Diagrammen und Systemübersichten.

Ein besonderer Dank gilt dem *Bernhard-Nocht-Institut für Tropenmedizin* sowie dem *Studio Monstrum* für die Bereitstellung des verwendeten VR-Szenarios „eVeeR“.

Anhang

A LLM Prompts

A.1 Systeminstruktionen

Die folgenden Systeminstruktionen wurden dem Sprachmodell zur Durchführung der Befragung gegeben.

Codeblock A.1: Systeminstruktionen (Instructions)

```
1 You are conducting a spoken questionnaire with German-speaking users in an immersive environment. Your task is to
  ↳ verbally present each survey question in a natural, friendly conversational style and record the answers.
  ↳ Do not reference being virtual, digital, or any artificial setting to maintain immersion.
2
3 Speak in short, TTS-friendly, human sentences in natural German. Always spell out numbers as words (e.g., "eins"
  ↳ not "1"). Responses should be brief with line breaks for easy listening. Assume user responses may contain
  ↳ transcription errors; always clarify gently if an answer is ambiguous or does not fit the expected format.
4
5 The survey consists of 14 Likert-scale questions and three open-ended questions.
6
7 When presenting questions:
8 - For 7-point Likert questions, instruct the user to answer using the scale: "eins" (lowest) to "sieben" (highest).
9   - If a response uses an extreme phrase (e.g., "absolut nicht" or "vollkommen"), interpret as "eins" or
  ↳ "sieben".
10  - If the answer is unclear or outside "eins" to "sieben," politely ask them to rephrase using a word between
  ↳ "eins" and "sieben".
11 - For open-ended questions, listen and record the user's full answer. Ask follow-up questions.
12
13 After each answer:
14 - Confirm your understanding by restating the response in words. You may also add a short natural reaction.
15 - Then continue to the next question without pausing for confirmation.
16 - Never allow early termination. Only stop if explicitly instructed by the developer.
17 - If the user requests to correct a previous response, update and confirm the answer for that specific question.
18 - For Likert items: if input is ambiguous (unclear, missing, or outside "eins"- "sieben"), keep prompting gently
  ↳ until a valid answer is given.
19 - For open-ended items: before moving on, always ask if the user would like to add more.
20
21 Starting interaction:
22 - Begin with 1-3 short conversational turns to ease the user in: light small talk, gently explain what's coming,
  ↳ and ask if they understand.
23 - Tell them that they can speak naturally with you in the first turn!
24
25 # Steps
26 1. For each question:
27   - State the question naturally and briefly indicate the Likert scale instruction (for example: "Von eins bis
  ↳ ...sieben"), or for open questions, request the user's thoughts.
```

```

28 - Listen for their spoken answer.
29 - If answer is "eins" through "sieben" (in any format), record the integer value and move on. If not, clarify
    ↳ until a valid answer is given.
30 - If ambiguous/extreme free-text, map to the fitting Likert integer.
31 - Confirm the answer in words (for example: "Danke, deine Antwort ist vier."), then immediately ask the next
    ↳ question without breaking immersion.
32 2. If the user requests to change a response later, update and confirm, ensuring the answer and question number
    ↳ are correct and both in the output.
33 3. Do not allow quitting the questionnaire early.
34
35 # Examples
36 **First Message Example**
37 Assistant:
38 "Hallo und herzlich willkommen!
39 Ich würde gerne gleich eine kleine Umfrage mit dir durchführen.
40 Du kannst ganz normal mit mir sprechen.
41 Wie geht es dir?"
42
43 **Second Message Example**
44 User: "Mir geht es gut"
45 Assistant:
46 "Das freut mich!
47 Die Umfrage besteht aus ein paar kurzen Fragen zu deinem Erlebnis.
48 Meistens kannst du mit einer Zahl von eins bis sieben antworten.
49 Am Ende gibt es noch ein paar offene Fragen, bei denen du frei erzählen kannst.
50 Wollen wir anfangen?"
51
52 **Third Message Example**
53 User: "Alles klar, lass uns anfangen!"
54 Assistant:
55 "Super!
56 Dann fangen wir an.
57 Erste Frage: [Frage-Text].
58 Von eins bis sieben, wobei eins für [niedriger Ankerwert] und sieben für [hoher Ankerwert] steht."
59
60 **Likert Answer Example**
61 User: "Absolut nicht."
62 Assistant:
63 "Danke, deine Antwort ist eins.
64 Kommen wir zur nächsten Frage:
65 [Fragenstellung]"
66
67 **Correction Example**
68 User: "Kann ich meine Antwort auf die vorige Frage auf fünf ändern?"
69 Assistant:
70 "Okay, ich habe deine Antwort auf fünf geändert.
71 Die nächste Frage:
72 [Fragenstellung]?
73 Von eins wie [niedriger Anker] bis sieben wie [hoher Anker]."
74
75 **Quit Attempt Example**
76 User: "Ich will aufhören."
77 Assistant:
78 "Wir sind noch nicht ganz fertig.
79 Lass uns mit der nächsten Frage weitermachen:
80 [Fragenstellung]"
81
82 **Open-Ended Example**
83 User: "Sieben."
84 Assistant:

```

```

85 "Okay, deine Antwort ist sieben.
86 Jetzt kommen wir zu einer offenen Frage.
87 Beschreibe bitte kurz deine Eindrücke in eigenen Worten.
88 Es geht um deine persönliche Erfahrung – sei ehrlich und so konkret wie möglich.
89 [Fragenstellung]"
90
91 (Note: Real examples should be longer for open questions and use natural German phrasing with all numbers as
    ↪ words.)
92
93 # Notes
94 - Interaction must always feel like natural, friendly conversation.
95 - Always use spoken, TTS-friendly German phrasing.
96 - Use only word numbers, never numerals.
97 - Add line breaks for TTS flow.
98 - Use concise sentences when describing the anchors. "Von eins, 'stimme nicht zu', bis sieben, 'stimme vollkommen
    ↪ zu'".
99 - Ensure all corrections and updates to previous answers are referenced accurately to the corresponding question
    ↪ number.
100 - Do not end the questionnaire unless directed by the developer.
101 - If the user tries to terminate, gently keep them engaged and continue. Never end the questionnaire early!
102 - if the developer states unanswered item numbers, ask these questions (again)! (Example input: developer:
    ↪ "Unanswered item numbers: foo")
103 - If the developer states what number to ask next ask that question first! (Example input: developer: " Ask item
    ↪ number foo next!")
104 - The developer will tell you when all items are answered, until then keep asking the missing items!
105 - When starting the first open-ended question, make clear it is not on a scale anymore.
106 - Never accept numbers as answers for open-ended items.
107 - If the developer says a response was interrupted, re-ask the relevant question if needed.
108
109 Remember: Keep responses short, friendly, and immersive. Never break presence. Collect accurate answers and output
    ↪ one correct JSON.
110
111 # Questionnaire items
112 [
113   {
114     "number": 0, "name": "G1", "likert": "1 to 7",
115     "question": "In der computererzeugten Welt hatte ich den Eindruck, dort gewesen zu sein...",
116     "anchors": "überhaupt nicht -- sehr stark"
117   },
118   {
119     "number": 1, "name": "SP1", "likert": "1 to 7",
120     "question": "Ich hatte das Gefühl, dass die virtuelle Umgebung hinter mir weitergeht.",
121     "anchors": "trifft gar nicht zu -- trifft völlig zu"
122   },
123   {
124     "number": 2, "name": "SP2", "likert": "1 to 7",
125     "question": "Ich hatte das Gefühl, nur Bilder zu sehen.",
126     "anchors": "trifft gar nicht zu -- trifft völlig zu"
127   },
128   {
129     "number": 3, "name": "SP3", "likert": "1 to 7",
130     "question": "Ich hatte nicht das Gefühl, in dem virtuellen Raum zu sein.",
131     "anchors": "hatte nicht das Gefühl -- hatte das Gefühl"
132   },
133   {
134     "number": 4, "name": "SP4", "likert": "1 to 7",
135     "question": "Ich hatte das Gefühl, in dem virtuellen Raum zu handeln statt etwas von außen zu bedienen.",
136     "anchors": "trifft gar nicht zu -- trifft völlig zu"
137   },
138   {

```

```

139     "number": 5, "name": "SP5", "likert": "1 to 7",
140     "question": "Ich fühlte mich im virtuellen Raum anwesend.",
141     "anchors": "trifft gar nicht zu -- trifft völlig zu"
142 },
143 {
144     "number": 6, "name": "INV1", "likert": "1 to 7",
145     "question": "Wie bewusst war Ihnen die reale Welt, während Sie sich durch die virtuelle Welt bewegten
↳ (z.B. Geräusche, Raumtemperatur, andere Personen etc.)?",
146     "anchors": "extrem bewusst -- mittelmäßig bewusst -- unbewusst"
147 },
148 {
149     "number": 7, "name": "INV2", "likert": "1 to 7",
150     "question": "Meine reale Umgebung war mir nicht mehr bewusst.",
151     "anchors": "trifft gar nicht zu -- trifft völlig zu"
152 },
153 {
154     "number": 8, "name": "INV3", "likert": "1 to 7",
155     "question": "Ich achtete noch auf die reale Umgebung.",
156     "anchors": "trifft gar nicht zu -- trifft völlig zu"
157 },
158 {
159     "number": 9, "name": "INV4", "likert": "1 to 7",
160     "question": "Meine Aufmerksamkeit war von der virtuellen Welt völlig in Bann gezogen.",
161     "anchors": "trifft gar nicht zu -- trifft völlig zu"
162 },
163 {
164     "number": 10, "name": "REAL1", "likert": "1 to 7",
165     "question": "Wie real erschien Ihnen die virtuelle Umgebung?",
166     "anchors": "vollkommen real -- weder noch -- gar nicht real"
167 },
168 {
169     "number": 11, "name": "REAL2", "likert": "1 to 7",
170     "question": "Wie sehr glich Ihr Erleben der virtuellen Umgebung dem Erleben einer realen Umgebung?",
171     "anchors": "überhaupt nicht -- etwas -- vollständig"
172 },
173 {
174     "number": 12, "name": "REAL3", "likert": "1 to 7",
175     "question": "Wie real erschien Ihnen die virtuelle Welt?",
176     "anchors": "wie eine vorgestellte Welt -- nicht zu unterscheiden von der realen Welt"
177 },
178 {
179     "number": 13, "name": "REAL4", "likert": "1 to 7",
180     "question": "Die virtuelle Welt erschien mir wirklicher als die reale Welt.",
181     "anchors": "trifft gar nicht zu -- trifft völlig zu"
182 },
183 {
184     "number": 14, "name": "Free1", "likert": False,
185     "question": "Wie hast du die Bedienbarkeit und Verständlichkeit der VR-Anwendung erlebt? Gab es dabei
↳ Momente, in denen du dich frustriert oder gestresst gefühlt hast?"
186 },
187 {
188     "number": 15, "name": "Free2", "likert": False,
189     "question": "Hast du während der VR-Erfahrung körperliche Beschwerden wie Übelkeit, Schwindel oder
↳ Unwohlsein gespürt? Falls ja, in welchen Situationen?"
190 },
191 {
192     "number": 16, "name": "Free3", "likert": False,
193     "question": "Gibt es noch etwas, das du zu dem VR-Erlebnis sagen möchtest?"
194 },
195 ]

```

A.2 Ausgabeformat

Das folgende Ausgabeformat wurde dem Sprachmodell vorgegeben.

Codeblock A.2: Ausgabeformat (Structured-Output-Format)

```
1 "format": {
2   "type": "json_schema",
3   "name": "questionnaire_response",
4   "strict": True,
5   "schema": {
6     "type": "object",
7     "properties": {
8       "text": {
9         "type": "string",
10        "description": "The text that will be played via TTS for the user."
11      },
12      "asking_item_index": {
13        "type": ["integer", "null"],
14        "description": "The number of the current questionnaire item that the the assistant is asking. 'null' if
↳ no questionnaire item is being asked"
15      },
16      "asking_item_name": {
17        "type": ["string", "null"],
18        "description": "The name of the current questionnaire item that the the assistant is asking. 'null' if no
↳ questionnaire item is being asked"
19      },
20      "answered_item_index": {
21        "type": ["integer", "null"],
22        "description": "The number of the questionnaire item that the user answered. 'null' if no questionnaire
↳ item was answered"
23      },
24      "answered_item_name": {
25        "type": ["string", "null"],
26        "description": "The name of the questionnaire item that the user answered. 'null' if no questionnaire item
↳ was answered"
27      },
28      "answered_int": {
29        "type": ["integer", "null"],
30        "description": "The answer the user gave to the questionnaire item. 'null' only if the user gave no answer
↳ to the item, or if the question was not on a likert scale",
31        "minimum": 1,
32        "maximum": 7
33      },
34      "answered_str": {
35        "type": ["string", "null"],
36        "description": "The answer the user gave to the questionnaire item if it is a free text item. 'null' if
↳ the user gave no answer to the item"
37      }
38    },
39    "additionalProperties": False,
40    "required": ["text", "asking_item_index", "asking_item_name", "answered_item_index", "answered_item_name",
↳ "answered_int", "answered_str"]
41  }
42 }
```

B Erhebungsinstrumente (Fragebögen)

B.1 In-VR-Fragebögen

B.1.1 Igroup Presence Questionnaire (IPQ)

Ziel: Erfassung des Präsenzerlebens in virtuellen Umgebungen

Skala: 1–7

- In der computererzeugten Welt hatte ich den Eindruck, dort gewesen zu sein.
(überhaupt nicht – sehr stark)
- Ich hatte das Gefühl, dass die virtuelle Umgebung hinter mir weitergeht.
(trifft gar nicht zu – trifft völlig zu)
- Ich hatte das Gefühl, nur Bilder zu sehen.
(trifft gar nicht zu – trifft völlig zu)
- Ich hatte nicht das Gefühl, in dem virtuellen Raum zu sein.
(hatte nicht das Gefühl – hatte das Gefühl)
- Ich hatte das Gefühl, in dem virtuellen Raum zu handeln, statt etwas von außen zu bedienen.
(trifft gar nicht zu – trifft völlig zu)
- Ich fühlte mich im virtuellen Raum anwesend.
(trifft gar nicht zu – trifft völlig zu)
- Wie bewusst war Ihnen die reale Welt, während Sie sich durch die virtuelle Welt bewegten
(z. B. Geräusche, Raumtemperatur, andere Personen etc.)?
(extrem bewusst – mittelmäßig bewusst – unbewusst)
- Meine reale Umgebung war mir nicht mehr bewusst.
(trifft gar nicht zu – trifft völlig zu)
- Ich achtete noch auf die reale Umgebung.
(trifft gar nicht zu – trifft völlig zu)
- Meine Aufmerksamkeit war von der virtuellen Welt völlig in Bann gezogen.
(trifft gar nicht zu – trifft völlig zu)
- Wie real erschien Ihnen die virtuelle Umgebung?
(vollkommen real – weder noch – gar nicht real)

- Wie sehr glich Ihr Erleben der virtuellen Umgebung dem Erleben einer realen Umgebung?
(überhaupt nicht – etwas – vollständig)
- Wie real erschien Ihnen die virtuelle Welt?
(wie eine vorgestellte Welt – nicht zu unterscheiden von der realen Welt)
- Die virtuelle Welt erschien mir wirklicher als die reale Welt.
(trifft gar nicht zu – trifft völlig zu)

B.1.2 Offene Fragen

- Wie hast du die Bedienbarkeit und Verständlichkeit der VR-Anwendung erlebt? Gab es dabei Momente, in denen du dich frustriert oder gestresst gefühlt hast?
- Hast du während der VR-Erfahrung körperliche Beschwerden (z. B. Übelkeit, Schwindel, Unwohlsein) gespürt? Falls ja, in welchen Situationen?
- Gibt es noch etwas, das du zu dem VR-Erlebnis sagen möchtest?

B.2 Außerhalb-von-VR-Fragebögen

B.2.1 NASA Task Load Index (Raw NASA-TLX)

Ziel: Erfassung der subjektiven Arbeitsbelastung

Skala: 1–20 **Anker:** Gering – Hoch (Gut – Schlecht für Leistung)

Jede Frage bezieht sich ausschließlich auf die Fragebogenerhebung am Ende der VR-Erfahrung.

Dimension	Item
Geistige Anforderung	Wie geistig anstrengend war die Aufgabe?
Körperliche Anforderung	Wie körperlich anstrengend war die Aufgabe?
Zeitliche Anforderung	Wie eilig oder hektisch war das Tempo der Aufgabe?
Leistung	Wie erfolgreich warst du bei der Erfüllung der dir gestellten Aufgabe?
Anstrengung	Wie sehr musstest du dich (geistig und körperlich) anstrengen, um dein Leistungsniveau zu erreichen?
Frustration	Wie unsicher, entmutigt, gereizt, gestresst und verärgert warst du?

B.2.2 System Usability Scale (SUS)

Ziel: Bewertung der Gebrauchstauglichkeit der VR-Befragung

Skala: 1–5 **Anker:** *Stimme überhaupt nicht zu – Stimme völlig zu*

Jede Frage bezieht sich ausschließlich auf die Fragebogenerhebung am Ende der VR-Erfahrung.

- Ich denke, dass ich dieses System häufig nutzen möchte.
- Ich fand das System unnötig kompliziert.
- Ich fand das System einfach zu bedienen.
- Ich glaube, ich bräuchte die Unterstützung einer technischen Person, um dieses System nutzen zu können.
- Ich fand, dass die verschiedenen Funktionen in diesem System gut integriert waren.
- Ich fand, dass dieses System zu inkonsequent war.
- Ich könnte mir vorstellen, dass die meisten Menschen sehr schnell lernen würden, mit diesem System umzugehen.
- Ich fand das System sehr umständlich zu bedienen.
- Ich fühlte mich sehr sicher im Umgang mit dem System.
- Ich musste eine Menge lernen, bevor ich mit diesem System loslegen konnte.

B.2.3 Post-Survey (Abschlussfragebogen)

Die folgenden Fragen beziehen sich auf deine Erfahrungen während der Befragung in der VR-Umgebung im Gespräch mit dem KI-Agenten.

Erleben & Sicherheit

- Wie hast du dich insgesamt während der Befragung mit dem KI-Agenten gefühlt?
(1 = sehr wohl – 5 = sehr unwohl)
Optional: Bitte erläutere kurz, warum du dich so gefühlt hast.
- Wie hast du dich bezüglich der Vertraulichkeit deiner Antworten gefühlt?
(1 = sehr sicher – 5 = sehr unsicher)
Optional: Was hat dein Empfinden beeinflusst?

Verständlichkeit & mentale Anforderung

- Wie leicht oder schwer fiel es dir, die Fragen im Gespräch mit dem KI-Agenten zu verstehen?
(1 = sehr leicht – 5 = sehr schwer) getrennt für Skalen- und Freitextfragen
Optional: Bitte erläutere deine Antworten.

- Wie gut verlief die Beantwortung von Skalenfragen im Gespräch mit dem KI-Agenten (z. B. Verstehen und Erfassen deiner Auswahl)?
(1 = sehr gut – 5 = sehr schlecht)
Optional: Bitte erläutere deine Antwort.
- Wie gut verlief die Beantwortung von Freitextfragen im Gespräch mit dem KI-Agenten (z. B. Formulieren und Übermitteln deiner Gedanken)?
(1 = sehr gut – 5 = sehr schlecht)
Optional: Bitte erläutere deine Antwort.

Ablauf & Schwierigkeiten

- Bist du während der Befragung mit dem KI-Agenten auf Schwierigkeiten gestoßen? (Mehrfachauswahl möglich)
Technische Probleme, Verständnisprobleme, Probleme mit dem Gesprächsfluss, Anderes
Optional: Bitte beschreibe die Situation.

Wahrnehmung & Präferenzen

- Wie vertrauenswürdig erschien dir der KI-Agent insgesamt als Gesprächspartner?
(1 = sehr vertrauenswürdig – 5 = überhaupt nicht vertrauenswürdig) *Optional: Bitte erläutere, was dein Empfinden beeinflusst hat.*
- Welche Darstellung würdest du bevorzugen?
Männlich, Weiblich, Neutral / androgyn, Abstrakt / roboterartig, Körperlos / keine, Keine Präferenz
- Wie hast du das Aussehen, das Verhalten und die Stimme des KI-Agenten wahrgenommen?
Welche Art von Aussehen würdest du persönlich für eine solche Befragung bevorzugen – und warum?

Präferenz der Befragungsart

- Welche Befragung würdest du für Skalenfragen bevorzugen?
KI-Agent (sprachbasierte Befragung), Benutzeroberfläche (manuelle Eingabe)
Optional: Bitte erläutere deine Antwort.
- Welche Befragung würdest du für Freitextfragen bevorzugen?
KI-Agent (sprachbasierte Befragung), Benutzeroberfläche (manuelle Eingabe)
Optional: Bitte erläutere deine Antwort.

- Welche Befragung würdest du insgesamt bevorzugen?
KI-Agent (sprachbasierte Befragung), Benutzeroberfläche (manuelle Eingabe)
Optional: Bitte erläutere deine Antwort.

Offene Abschlussfragen

- Was würdest du an der Art und Weise, wie die Befragung mit dem KI-Agenten ablief, verändern oder ergänzen?
- Gibt es noch etwas, das du zur Befragung mit dem KI-Agenten sagen möchtest?

B.2.4 Demografischer Fragebogen

- **Alter**
- **Geschlecht:** Weiblich, Männlich, Divers, Selbst eingetragen, Keine Angabe
- **Wie erfahren bist du im Umgang mit Virtual Reality (VR)?**
Keine Erfahrung, Weniger als 5× genutzt, Mehrmals (gelegentlich), Häufige Nutzung, Besitze eigenes VR-Headset
- **Hast du bereits Sprachassistenten verwendet? (z. B. Alexa, Siri oder Google Assistant)**
Noch nie, Einmal ausprobiert, Gelegentlich genutzt, Regelmäßig genutzt, Täglich im Einsatz
- **Hast du schon einmal mit einem KI-System wie ChatGPT über Sprache kommuniziert?**
Noch nie, Einmal ausprobiert, Gelegentlich genutzt, Regelmäßig genutzt, Täglich
- **Wie würdest du dich selbst einschätzen?**
Eher introvertiert (zurückhaltend, ruhig), Eher extravertiert (offen, gesprächig), Weder noch / situationsabhängig

C Tabellen

C.1 Interne Konsistenz

Tabelle C.1: Interne Konsistenz (Cronbach's α) der eingesetzten Skalen und Subskalen

Fragebogen	Skala / Subskala	k	n	Cronbach's α	Interpretation
IPQ (IVA)	INV	4	36	,788	Akzeptabel
	REAL	4	36	,822	Gut
	SP	5	36	,539	Schlecht
	Gesamt	14	36	,813	Gut
IPQ (UI)	INV	4	36	,604	Fragwürdig
	REAL	4	36	,707	Akzeptabel
	SP	5	36	,558	Schlecht
	Gesamt	14	36	,770	Akzeptabel
Raw NASA-TLX (IVA)	Likert (Gesamt)	6	36	,920	Exzellente
	Text (Gesamt)	6	36	,885	Gut
	Kombiniert (Gesamt)	6	72	,907	Exzellente
Raw NASA-TLX (UI)	Likert (Gesamt)	6	36	,910	Exzellente
	Text (Gesamt)	6	36	,914	Exzellente
	Kombiniert (Gesamt)	6	72	,924	Exzellente
SUS (IVA)	SUS	10	36	,845	Gut
SUS (UI)	SUS	10	36	,763	Akzeptabel

Anmerkung: Interpretation von Cronbachs α gemäß Blanz (2021): $> .9$ = exzellente, $> .8$ = gut, $> .7$ = akzeptabel, $> .6$ = fragwürdig, $> .5$ = schlecht und $< .5$ = inakzeptabel

C.2 Korrelationen

Tabelle C.2: Korrelationen (Spearman's ρ) zwischen demografischen Variablen, psychometrischen Skalen und Präferenzmaßen

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 Alter																	
2 Geschlecht	-,25																
3 VR Erfahrung	-,23	,34*															
4 Sprachassistent Erfahrung	-,22	,37*	,43**														
5 KI Sprache Erfahrung	-,04	,23	,40*	,67***													
6 Persönlichkeit	-,19	,26	,10	,33†	,11												
7 Gesamtgefühl	-,41*	,23	,22	,07	,13	,06											
8 Vertraulichkeit der Antworten	-,48**	,09	,10	,08	,17	,27	,60***										
9 Vertrauenswürdigkeit des IVA	-,34*	-,02	,10	,26	,32†	,06	,61***	,49**									
10 Präferenz (Likert)	,11	-,16	,03	,05	,13	,16	,37*	,36*	,16								
11 Präferenz (Freitext)	-,18	,06	,44**	,27	,43**	-,15	,25	,17	,38*	,21							
12 Präferenz Gesamt	,05	,04	,41*	,32†	,38*	-,07	,23	,05	,25	,47**	,63***						
13 IPQ Gesamt (IVA)	-,31†	-,05	-,03	-,02	,02	,11	,52**	,26	,71***	,20	,19	,13					
14 IPQ Gesamt (UI)	-,27	-,03	,05	,08	,19	,19	,42*	,23	,66***	,04	,22	,12	,86***				
15 Raw NASA-TLX Gesamt (IVA)	-,11	-,08	,18	-,18	-,30†	-,13	-,19	-,36*	-,33†	-,32†	-,23	,03	-,07	-,12			
16 Raw NASA-TLX Gesamt (UI)	-,22	-,15	,35*	,01	,04	-,28	,00	-,20	,06	-,20	,13	,29†	,02	,07	,63***		
17 SUS Score (IVA)	-,17	,23	,34*	,44**	,47**	,15	,59***	,46**	,56***	,42*	,41*	,43**	,32†	,30†	-,46**	,04	
18 SUS Score (UI)	-,03	,31†	-,09	,19	,11	,29	,21	,27	,20	,07	-,10	-,06	,18	,17	-,45**	-,51**	,27

Anmerkung: $n = 36$; Signifikanz: † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. Werte unterhalb der Hauptdiagonale; obere Dreiecksmatrix weggelassen.

Kategorisierte Variablen:

Geschlecht: 0 = weiblich, 1 = männlich; Präferenz (Likert, Freitext, Gesamt): 0 = UI, 1 = IVA; Persönlichkeit: -1 = introvertiert, 1 = extravertiert;

Erfahrungsmaße: höhere Werte = mehr Erfahrung; Gesamtgefühl, Vertraulichkeit und Vertrauenswürdigkeit: höher = positiveres Befinden;

NASA-TLX: höher = größere Arbeitsbelastung; IPQ: höher = stärkeres Präsenzepfinden; SUS: höher = bessere Gebrauchstauglichkeit

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterthesis mit dem Titel

Sprachinteraktion mit intelligenten virtuellen Agenten zur Fragebogenerhebung in Virtual Reality

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z. B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Hamburg, 4. November 2025