ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Informatik & Security

fh st.pölten

WE ARE EUDRES EUROPEAN UNIVERSITY

Einfluss humanoider Roboter im Vergleich zu Web-Chatbots mit Retrieval-Augmented Generation (RAG) auf Akzeptanz und Benutzerfreundlichkeit im Bildungsbereich

Masterarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in (Dipl.-Ing.)

eingereicht von

Marcel Dielacher, BSc.
11906513

im Rahmen des
Studiengangs Data Intelligence an der Fachhochschule St. Pölten

Betreuung
Betreuer/Betreuerin: Thomas Delissen, MSc.

Abstract

The rapid digitalization of education is opening new possibilities and fundamentally transforming established teaching and learning processes. Artificial intelligence (AI) now plays an increasingly prominent role, especially through interactive technologies such as humanoid robots and retrieval-augmented generation (RAG) web chatbots. While humanoid robots enable natural communication thanks to their physical presence and multimodal interaction, RAG-based web chatbots offer flexible, data-driven support. Despite their growing use, a systematic comparison of their acceptance and usability in educational settings is still lacking.

This master's thesis therefore compares humanoid robots and RAG-enhanced web chatbots with regard to user experience, acceptance and potential effects on learning behavior. A total of 20 participants were included in the study. They first completed a pretest, then interacted either with the humanoid robot or with the web chatbot, followed by online- questionnaire and a post-test. All data were collected empirically via a structured questionnaire, combining quantitative and qualitative measures. The analyses employed the UTAUT and SUS models, additionally, a custom RAG pipeline was implemented with LlamaIndex to optimize information delivery.

NAO achieved very high pretest results, but these declined drastically after the interaction, whereas the web chatbot showed very stable performance in both the pretest and posttest. In terms of acceptance, the chatbot achieved slightly higher values (UTAUT = 3.68) than NAO (3.57). Regarding usability, NAO scored higher (SUS = 3.01) compared to the chatbot (2.78). Occasional AI users rated both systems most positively, while gender differences were marginal. Overall, the robot demonstrated strength in usability and incremental learning improvement, whereas the chatbot was favored for acceptance and adaptability.

The findings illustrate that both AI-based systems possess distinct strengths, the robot excels in usability, whereas the web chatbot enjoys greater acceptance but with more stable performance. Consequently, the results provide both a scientific contribution to the study of AI-supported educational technologies and practical guidance on which technology is most advantageous for specific learning environments and target groups.

Zusammenfassung

Die rasante Digitalisierung der Bildung eröffnet neue Möglichkeiten und transformiert Lehr- und Lernprozesse grundlegend. Künstliche Intelligenz (KI) spielt dabei eine zunehmend wichtige Rolle, insbesondere durch interaktive Technologien wie humanoide Roboter, Retrieval-Augmented-Generation-(RAG) und Web-Chatbots. Während humanoide Roboter durch ihre physische Präsenz und multimodale Interaktion eine besonders natürliche Kommunikation ermöglichen, bieten RAG-basierte Web-Chatbots flexible, datengetriebene Unterstützung. Trotz ihrer wachsenden Verbreitung fehlt bislang ein systematischer Vergleich hinsichtlich Akzeptanz und Nutzbarkeit im Bildungskontext.

Diese Masterarbeit vergleicht daher humanoide Roboter und RAG-gestützte Web-Chatbots im Hinblick auf Benutzerfreundlichkeit, Akzeptanz und potenzielle Effekte auf das Lernverhalten. Insgesamt nahmen 20 Personen an der Studie teil. Sie absolvierten zunächst einen Pretest, interagierten anschließend entweder mit dem humanoiden Roboter oder mit dem Web-Chatbot, danach wurde ein Fragebogen ausgefüllt und schlossen mit einem Posttest ab. Die Datenerhebung erfolgte empirisch über einen strukturierten Fragebogen, der quantitative und qualitative Maße kombinierte. Zur Analyse wurden die Modelle UTAUT und SUS herangezogen, zudem wurde mit LlamaIndex eine RAG-Pipeline implementiert, um die Informationsbereitstellung zu optimieren.

NAO erzielte sehr hohe Ergebnisse im Pretest, diese nahmen jedoch nach der Interaktion drastisch ab, wohingegen der Web-Chatbot sowohl im Pre- als auch im Posttest eine sehr stabile Leistung zeigte. In Bezug auf die Akzeptanz erzielte der Chatbot leicht höhere Werte (UTAUT = 3,68) als NAO (3,57). Hinsichtlich der Usability schnitt NAO besser ab (SUS = 3,01) als der Chatbot (2,78). Gelegentliche KI-Nutzer*innen bewerteten beide Systeme am positivsten, während Geschlechterunterschiede nur marginal ausfielen. Insgesamt erwies sich der Roboter als stark in Bezug auf Benutzerfreundlichkeit und inkrementelle Lerngewinne, während der Chatbot in Akzeptanz und Anpassungsfähigkeit bevorzugt wurde.

Die Befunde verdeutlichen, dass beide KI-gestützten Systeme über spezifische Stärken verfügen. Der Roboter überzeugt durch hohe Usability, während der Web-Chatbot eine größere Akzeptanz genießt, jedoch mit stabilerem Leistungsniveau. Damit leisten die Ergebnisse sowohl einen wissenschaftlichen Beitrag zur Untersuchung KI-gestützter Bildungstechnologien als auch praktische Orientierung, welche Technologie sich für bestimmte Lernumgebungen und Zielgruppen am besten eignet.

Vorwort

Die Digitalisierung beeinflusst nicht nur unsere Arbeitswelt, sondern auch, wie wir lernen, lehren und Wissen verfügbar machen. Besonders im Bildungsbereich zeigt sich, wie tiefgreifend technologische Innovationen traditionelle Strukturen hinterfragen und neue Chancen schaffen können. Die Idee zur vorliegenden Masterarbeit entstand aus meinem persönlichen Interesse an der Frage, wie Bildung mithilfe von Künstlicher Intelligenz (KI) zugänglicher, individueller und unabhängiger von klassischen Bildungsstrukturen gestaltet werden kann.

Ich bin überzeugt davon, dass Bildung ein Grundrecht ist, das allen Menschen offenstehen sollte, unabhängig von Wohnort, Herkunft oder verfügbaren Ressourcen. Gerade in bildungsfernen Regionen oder in Gesellschaften, die mit einem zunehmenden Mangel an Lehrkräften konfrontiert sind, braucht es innovative Ansätze, um qualitativ hochwertige Bildungsangebote sicherzustellen. KI-gestützte Systeme wie Web-Chatbots oder humanoide Roboter können hier eine bedeutende Rolle spielen. Sie sind skalierbar, flexibel einsetzbar und potenziell in der Lage, Lernende individuell zu unterstützen, auch ohne direkte Betreuung durch Lehrpersonen.

Im Rahmen meines Masterstudiums Data Intelligence an der FH St. Pölten war es mir ein besonderes Anliegen, das Potenzial von KI nicht nur aus einer technischen Perspektive zu betrachten, sondern auch in Hinblick auf konkrete gesellschaftliche Herausforderungen zu analysieren. In dieser Arbeit steht deshalb die Frage im Mittelpunkt, wie sich zwei unterschiedliche Formen KI-gestützter Lernassistenz, ein webbasierter Chatbot mit Retrieval-Augmented Generation (RAG) und ein humanoider Roboter mit Retrieval-Augmented Generation, auf Akzeptanz und Benutzerfreundlichkeit auswirken. Beide Systeme verfügen über Dialogfähigkeiten, unterscheiden sich jedoch deutlich in ihrer Interaktionsform, sozialen Präsenz und Wahrnehmung durch die Nutzer*innen.

Die interdisziplinäre Herangehensweise dieser Arbeit zwischen Informatik, Psychologie, Pädagogik und Mensch-Roboter-Interaktion ermöglicht es, technologische Möglichkeiten nicht nur funktional zu betrachten, sondern stets im Hinblick auf ihre Wirkung auf den Menschen. Mein Ziel war es, einen Beitrag zur Frage zu leisten, wie wir KI nicht nur als Werkzeug, sondern als verantwortungsvoll gestalteten Bildungsbegleiter nutzen können insbesondere dort, wo menschliche Unterstützung fehlt oder nicht ausreicht.

Danksagung

An dieser Stelle möchte ich mich herzlich bei allen bedanken, die mich während der Entstehung dieser Masterarbeit unterstützt haben.

Mein besonderer Dank gilt Thomas Delissen, MSc, für die hervorragende Betreuung, die wertvollen Anregungen und das konstruktive Feedback während des gesamten Prozesses. Seine fachliche Expertise und Unterstützung haben maßgeblich zum Gelingen dieser Arbeit beigetragen.

Ebenso möchte ich mich bei dem gesamten Team der FH St. Pölten bedanken, das mir eine großartige Ausbildung ermöglicht und mich auf meinem akademischen Weg begleitet hat. Die inspirierende Lernumgebung und die erstklassige Lehre haben mich sowohl fachlich als auch persönlich enorm weiterentwickelt.

Ein großes Dankeschön gilt auch den Proband*innen, die bereit waren, an meiner Studie während einer Lehrveranstaltung teilzunehmen und mir damit wertvolle Einblicke ermöglicht haben. Ohne ihre Zeit und ihr Engagement wäre diese Arbeit in dieser Form nicht möglich gewesen.

Mein tiefster Dank geht an meine Freundin Sarah Lohninger, BSc. und an allen Freunden, die mich stets motiviert, unterstützt und in herausfordernden Zeiten an mich geglaubt haben. Ihre Geduld und ihr Verständnis haben mir sehr geholfen.

Nicht zuletzt danke ich meinen Eltern, die mir meine akademische Laufbahn finanziert und mich mit Rat und Tat unterstützt haben. Sie hatten immer ein offenes Ohr für mich und haben mich in jeder Hinsicht gefördert, ohne sie wäre mein akademischer Weg nicht möglich gewesen.

Eidesstaatliche Erklärung

Ich erkläre an Eides statt, dass

- ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.
- ich mich bei der Erstellung der Arbeit an die Standards guter wissenschaftlicher Praxis gemäß dem Leitfaden zum Wissenschaftlichen Arbeiten der FH St. Pölten gehalten habe.
- ich die vorliegende Arbeit an keiner Hochschule zur Beurteilung oder in irgendeiner Form als Prüfungsarbeit vorgelegt oder veröffentlicht habe.

Über den Einsatz von Hilfsmitteln der generativen Künstlichen Intelligenz wie Chatbots, Bildgeneratoren, Programmieranwendungen, Paraphrasier- oder Übersetzungstools erkläre ich, dass

☐ im Zuge dieser Arbeit kein Hilfsmittel der generativen Künstlichen Intelligenz zum Einsatz gekommen ist.

☒ ich Hilfsmittel der generativen Künstlichen Intelligenz verwendet habe, um die Arbeit Korrektur zu lesen.

☐ ich Hilfsmittel der generativen Künstlichen Intelligenz verwendet habe, um Teile des Inhalts der Arbeit zu erstellen. Ich versichere, dass ich jeden generierten Inhalt mit der Originalquelle zitiert habe. Das genutzte Hilfsmittel der generativen Künstlichen Intelligenz ist an entsprechenden Stellen ausgewiesen.

Durch den Leitfaden zum Wissenschaftlichen Arbeiten der FH St. Pölten bin ich mir über die Konsequenzen einer wahrheitswidrigen Erklärung bewusst.

Wien, 08.09.2025 Marcel Dielacher

Inhaltsverzeichnis

1 Einleitung

Die Digitalisierung hat in den letzten Jahrzehnten viele Lebensbereiche verändert und bringt insbesondere im Bildungssektor eine Vielzahl neuer Möglichkeiten mit sich [1]. Eine Entwicklung in diesem Zusammenhang ist die Integration von Künstlicher Intelligenz (KI) in Form von humanoiden Robotern und intelligenten Chatbots. Diese innovativen Ansätze bieten das Potenzial, Lernprozesse interaktiver, individueller und effizienter zu gestalten, indem sie personalisierte Unterstützung bieten und flexibel auf die Bedürfnisse der Lernenden eingehen [2]. Schon in den 1950er-Jahren, im Zusammenhang mit dem Turing-Test, entstand das grundlegende Konzept von Systemen, die in der Lage sind, mit Menschen zu kommunizieren, heute als Chatbots bekannt. Ein frühes Beispiel dafür war das Programm Eliza, das als eines der ersten seiner Art gilt [3].

Der Einsatz von KI im Bildungsbereich eröffnet neue Perspektiven für das Lernen, da diese Technologie es ermöglicht, herkömmliche Lehrmethoden zu erweitern oder auch zu revolutionieren [4]. Durch ihre Fähigkeit, natürliche Sprache zu erzeugen und interaktiv zu kommunizieren, können sie Lernprozesse bereichern und die Umsetzung des Sustainable Development Goals 4 unterstützen, vorausgesetzt, ihr Einsatz erfolgt reflektiert und verantwortungsvoll [5]. Student*innen oder Schüler*innen können beispielsweise mit einem humanoiden Roboter interagieren, der ihnen nicht nur auf Basis vorprogrammierter Algorithmen antwortet, sondern auch mithilfe von KI-Algorithmen individuelle Rückmeldungen geben kann. Alternativ können sie einen intelligenten Chatbot auf einem Laptop oder Tablet nutzen, um Fragen zu stellen, Inhalte nachzuschlagen oder sich gezielt prüfen zu lassen. In beiden Fällen entsteht ein interaktiver Lernprozess, der über die klassischen Unterrichtsformen hinausgeht. Besonders hervorzuheben ist, dass diese Technologien nicht nur reaktiv agieren, sondern auch proaktiv Fragen stellen können, um den Lernfortschritt zu überprüfen und gezielt Impulse zur Vertiefung des Wissens zu geben.

Diese Thematik gewinnt an Bedeutung, da weltweit Millionen von Menschen keinen oder nur eingeschränkten Zugang zu hochwertiger Bildung haben [6], [7], [8]. Die Ursachen hierfür sind vielfältig, da in vielen Ländern ein akuter Mangel an qualifizierten Lehrkräften herrscht, während gleichzeitig die Nachfrage nach Bildungsangeboten kontinuierlich steigt [9], [10]. Gerade in strukturschwachen Regionen oder in Ländern mit unzureichender Bildungsinfrastruktur könnte der Einsatz von KI-basierten Lernsystemen dazu beitragen, Bildungsangebote flexibler und für eine breitere Zielgruppe zugänglich zu machen [11]. Auch in hochentwickelten Ländern können diese Technologien zur Entlastung von Lehrkräften beitragen, indem sie Routineaufgaben übernehmen und es Lehrpersonal ermöglichen, sich stärker auf individuelle Förderung zu konzentrieren.

Ein Ansatz ist der Einsatz von Retrieval Augmented Generation (RAG), einer Technik, die Large Language Modelle (LLM) mit externen, spezifischen Wissensquellen kombiniert. Dadurch können generierte Informationen nicht nur auf allgemeinem Weltwissen der LLM's basieren, sondern gezielt mit relevanten, aktuellen und bereitgestellten Daten angereichert werden [12]. Dies ist in universitären Kontexten von Bedeutung, wo Studierende häufig mit komplexen, fachspezifischen Fragestellungen konfrontiert sind [13]. Während textbasierte webbasierte Chatbots bereits in zahlreichen Bereichen eingesetzt werden, bieten physische

Roboter mit sozialen Interaktionsfähigkeiten, wie der NAO-Roboter, eine neue Dimension der Mensch-Maschine-Kommunikation (MMK).

Gerade an dieser Stelle stellt sich die Frage, weshalb humanoide Roboter überhaupt in den Vergleich mit Chatbots aufgenommen werden sollten. Während der Nutzen von Chatbots und LLMs im Bildungskontext durch ihre ständige Verfügbarkeit, die schnelle Verarbeitung großer Datenmengen und die einfache Integration in digitale Lernumgebungen klar erkennbar ist, liegt die besondere Relevanz von Robotern in ihren physischen und sozialen Interaktionsmöglichkeiten. Sie verfügen über körperliche Präsenz, Gestik, Mimik und Sprachinteraktion, wodurch sie Lernende nicht nur kognitiv, sondern auch sozial und emotional ansprechen können. [14]

Gerade in Bildungskontexten, in denen soziale Präsenz, Motivation und Engagement zentrale Erfolgsfaktoren darstellen, ist es daher bedeutsam zu prüfen, ob Roboter in Kombination mit RAG-Technologien einen zusätzlichen Mehrwert gegenüber rein textbasierten Systemen liefern können. Während Chatbots ihre Stärken vor allem in der digitalen Flexibilität haben, könnten humanoide Roboter durch ihre physische und soziale Wirkung eine qualitativ andere Form der Unterstützung ermöglichen.

Damit ist der direkte Vergleich nicht nur ein technischer Test zweier Systeme, sondern eine zentrale Forschungsfrage:

Wie beeinflusst die Verwendung eines humanoiden Roboters im Vergleich zu einem Web-Chatbot mit RAG die Akzeptanz und die Benutzerfreundlichkeit im Bildungsbereich?

Diese Frage dient als Grundlage für die weitere Untersuchung und erlaubt es, sowohl technologische als auch nutzerzentrierte Perspektiven systematisch zu analysieren. Zur Beantwortung dieser übergeordneten Forschungsfrage werden folgende Hypothesen formuliert:

H1: Der NAO-Roboter wird von den Nutzer*innen als benutzerfreundlicher wahrgenommen als der webbasierte Chatbot.
H2: Die Leistungserwartung ist beim NAO-Roboter höher ausgeprägt als beim Web-Chatbot.
H3: Die Aufwandserwartung ist beim Web-Chatbot höher ausgeprägt (d.h. leichter zu bedienen) als beim NAO-Roboter.
H4: Die Sozialer Einfluss (wahrgenommene soziale Beeinflussung) ist beim NAO-Roboter stärker ausgeprägt als beim Web-Chatbot.
H5: Die erleichternden Bedingungen haben einen stärkeren Einfluss auf die Akzeptanz des NAO-Roboters als auf die des Web-Chatbots.
H6: Die Verhaltensabsicht zur weiteren Nutzung ist beim NAO-Roboter höher als beim Web-Chatbot.
H7: Der NAO erzielt höhere Akzeptanzwerte als der Web-Chatbot.
H8: Der Einsatz des NAO führt zu höheren Wissensvermittlung als der Einsatz eines webbasierten Chatbots.

Während in der wissenschaftlichen Literatur bereits zahlreiche Studien zur Akzeptanz und Benutzerfreundlichkeit einzelner Technologien existieren, fehlt bislang eine systematische Analyse, die humanoide Roboter und RAG-gestützte Web-Chatbots in einem direkten Vergleich betrachtet [15], [1], [16], [13], [17]. Die Arbeit zielt darauf ab, ein tiefergehendes

Verständnis für die Nutzungserfahrungen mit beiden Technologien zu gewinnen und mögliche Implikationen für deren Einsatz im schulischen sowie universitären Kontext abzuleiten.

Um die Forschungsfrage fundiert zu beantworten, wird eine Untersuchung durchgeführt, die qualitative Forschungsmethoden kombiniert. Für das Experiment wurde eine RAG-Pipeline mit LLama Index programmiert, um eine optimierte Retrieval-Funktionalität und eine verbesserte Bereitstellung von aktuellen und relevanten Informationen von Lehrmaterial zu gewährleisten. Die Studie umfasst mehrere methodische Ansätze, um eine umfassende Analyse der Benutzerakzeptanz und der wahrgenommenen Benutzerfreundlichkeit zu ermöglichen. Zentral für die Datenerhebung sind dabei Benutzerstudien, in denen Teilnehmende mit einem humanoiden Robotern oder mit einem RAG-gestützten Web-Chatbots interagieren. Dabei wird untersucht, wie unterschiedlich die Akzeptanz und die Benutzerfreundlichkeit der beiden Technologien ausfällt und welche spezifischen Vor- und Nachteile sie in der Lernumgebung mit sich bringen.

Zentral für die Datenerhebung sind dabei Benutzerstudien, in denen Teilnehmende in zwei Gruppen aufgeteilt werden, eine Gruppe interagiert mit dem webbasierten Chatbot, die andere mit dem humanoiden Roboter NAO. Beide Systeme nutzen dieselbe RAG-Pipeline mit gleichen funktionalen Fähigkeiten zur Informationsverarbeitung. Der Unterschied liegt in der Art der Interaktion, während der NAO-Roboter über Sprachsynthese sowie multimodale Ausdrucksformen wie leichter Gestik verfügt, besitzt der webbasierte Chatbot keine Sprachausgabe und interagiert rein textbasiert über eine Weboberfläche.

Die empirische Untersuchung stützt sich auf das Unified Theory of Acceptance and Use of Technology-Modell (UTAUT) [18] sowie auf die System Usability Scale (SUS) [19], um eine differenzierte Bewertung der Nutzerakzeptanz und der Benutzerfreundlichkeit vorzunehmen. Während das UTAUT-Modell insbesondere Faktoren wie Leistungserwartung (LE), Aufwandserwartung (AE), soziale Einflüsse (SE) und Nutzungskontext [18] berücksichtigt, dient das SUS-Modell zur standardisierten Erfassung der wahrgenommenen Gebrauchstauglichkeit der jeweiligen Technologien [19].

Zusätzlich zu diesen Faktoren wird in der Untersuchung ein Pre- und Post-Wissenstest durchgeführt, um mögliche Lernfortschritte der Teilnehmer*innen zu messen. Dieser Test dient als objektive Messgröße für die Effektivität der jeweiligen Technologie im Hinblick auf Wissensvermittlung und Lernerfolg. Es wird erwartet, dass humanoide Roboter durch ihre physische Präsenz und multimodale Interaktionsmöglichkeiten, etwa durch Gestik und der Sprachintonation, einen positiven Einfluss auf die Lernmotivation und das Engagement der Teilnehmenden haben. Gleichzeitig könnte sich zeigen, dass RAG-gestützte webbasierte Chatbots durch ihre hohe Verfügbarkeit, schnelle Reaktionszeiten und datenbasierte Präzision ebenso gut oder in bestimmten Kontexten vorteilhafter für den Lernprozess sind.

Bereits eine Studie von Neumann et al. [17] hat herausgefunden, dass die physische Präsenz eines Roboters das Engagement und die Motivation der Lernenden steigern kann, insbesondere in Szenarien, in denen direkte Interaktion und sozialer Austausch eine Rolle spielen. Andererseits bieten webbasierte Chatbots eine höhere Flexibilität, da sie orts- und zeitunabhängig genutzt werden können. Dies macht sie besonders für asynchrones Lernen in digitalen Bildungsformaten attraktiv. Studierende greifen zunehmend auf Large Language Models zurück, um ihre akademischen Aufgaben effizienter zu bewältigen [17]. Seit der Veröffentlichung von ChatGPT durch OpenAI Ende 2022 hat die Akzeptanz von Chatbots stark

zugenommen. Das zugrunde liegende LLM hat nicht nur dazu beigetragen, generative KI-Technologien in der breiten Öffentlichkeit zu etablieren, sondern auch grundlegende Veränderungen im Verständnis und in der Nutzung Künstlicher Generativer Intelligenz (KGI) in unterschiedlichsten Lebensbereichen angestoßen [2].

Die Ergebnisse dieser Studie sollen nicht nur einen wissenschaftlichen Beitrag zur aktuellen Forschungslage leisten, sondern auch praxisnahe Handlungsempfehlungen für die Implementierung von KI im Bildungsbereich liefern. Vor allem für Schulen, Universitäten und andere Bildungseinrichtungen kann diese Untersuchung wichtige Erkenntnisse darüber liefern, ob ein Web-Chatbot oder ein humanoider Roboter mit einem RAG-System mehr Aufmerksamkeit erhält.

Ein zentrales methodisches Limit liegt in der geringen Zahl direkter Vergleichsstudien zwischen humanoiden Robotern und RAG-gestützten Web-Chatbots im Bildungskontext. Während Roboterforschung oft auf soziale Präsenz und Motivation fokussiert, untersuchen Chatbot-Studien überwiegend textbasierte Systeme, RAG-Technologien werden nur selten systematisch evaluiert. Dadurch fehlen belastbare Referenzwerte mit vergleichbarem Design, insbesondere für UTAUT und SUS Dimensionen.

Diese Arbeit ist in mehrere thematische Abschnitte untergliedert, um die Komplexität des Untersuchungsgegenstandes systematisch zu erfassen. Zu Beginn wird der theoretische Hintergrund erläutert, insbesondere der aktuelle Stand der Technik im Bereich der Künstlichen Intelligenz in der Pädagogik, einschließlich Web-Chatbots, humanoider Roboter und verwandter Technologien. Anschließend liegt der Fokus auf dem durchgeführten Experiment, der eingesetzten Technologie sowie der verwendeten Methodik. Darauf folgen die Darstellung der Ergebnisse und eine ausführliche Diskussion. Abschließend wird das im Rahmen der Literaturrecherche angewandte PRISMA-Framework grafisch aufbereitet, um die Auswahl und Einbeziehung relevanter Studien transparent nachvollziehbar zu machen.

2 Theoretischer Hintergrund

Dieses Kapitel dieser Arbeit dient als grundlegende Einführung in die thematische Ausrichtung und Zielsetzung der Untersuchung. Aufbauend auf dem theoretischen Hintergrund werden hier die zentralen Dimensionen beschrieben, die für das Verständnis der weiteren Ausführungen von Relevanz sind. Ziel ist es, einen klaren Überblick über die Forschungsfrage, den Kontext sowie die methodische Herangehensweise zu geben und somit den inhaltlichen Rahmen für die Analyse abzustecken.

Zunächst wird der Stand der Technik beschrieben, danach wird die Bedeutung von Künstlicher Intelligenz im Bildungskontext eingeordnet. Dabei geht es um die Frage, wie KI-basierte Systeme, insbesondere Web-Chatbots und humanoide Roboter, in schulischen und hochschulischen Lernumgebungen eingesetzt werden können und welche Potenziale, aber auch Einschränkungen, sich daraus ergeben. Ergänzend wird auf die technologischen Grundlagen eingegangen, die für das Verständnis der späteren Analysen essenziell sind.

Ein weiterer Schwerpunkt liegt auf der Rolle von Retrieval-Augmented Generation (RAG) Pipelines und Large Language Models (LLMs), deren Funktionsweise sowie deren Anwendungsfelder im Bildungsbereich. Diese Technologien werden nicht nur technisch beschrieben, sondern auch hinsichtlich ihrer Bedeutung für Lernprozesse und pädagogische Interaktionen diskutiert.

Darüber hinaus werden interdisziplinäre Perspektiven aufgegriffen, die das Thema um psychologische und sozialwissenschaftliche Sichtweisen erweitern.

Ein zentrales Element dieses Kapitels ist schließlich die Vorstellung relevanter Modelle zur Bewertung von Technologieakzeptanz und Benutzerfreundlichkeit, das UTAUT-Modell sowie die System Usability Scale (SUS). Beide Ansätze bilden die methodische Grundlage für diese empirische Untersuchung und erlauben eine strukturierte Analyse von Nutzungswahrnehmungen und -erfahrungen.

Insgesamt soll dieses Kapitel eine klare Orientierung bieten, indem es die wesentlichen theoretischen, technologischen und methodischen Eckpfeiler darstellt. Es bildet damit die Brücke zwischen dem allgemeinen Forschungsstand und der empirischen Untersuchung und schafft die notwendige Grundlage für eine kritische Auseinandersetzung mit Chancen und Herausforderungen des KI-Einsatzes im Bildungskontext.

2.1 Stand der Forschung

Mit dem rasanten Aufstieg von LLMs wie ChatGPT sind zunehmend Anwendungen im Bildungsbereich entstanden, die das Potenzial haben, Lernprozesse zu transformieren. Insbesondere der Einsatz von LLM-basierten Chatbots als virtuelle Tutoren verspricht neue Möglichkeiten der individuellen Förderung. Aktuelle Forschung wie die von Lieb et al. [1] zeigt, dass ein gezieltes Prompt-Engineering entscheidend für die Wirksamkeit solcher Chatbots ist. In ihrer Studie entwickelten die Autor*innen den Chatbot NewtBot, der auf GPT-3.5 basiert und

speziell für den Physikunterricht an weiterführenden Schulen konzipiert wurde. Im Fokus steht die interne Anpassung des Modells über sogenannte Systemprompts, Anweisungen, die das Verhalten des Chatbots steuern, ohne dass die Nutzer*innen diese sehen. Drei verschiedene Konfigurationen wurden getestet, ein generisches Basismodell und ein tutororientiertes Modell mit didaktischer Gesprächsführung sowie ein aufgabenbezogenes Feedback-Modell. Die Evaluation mit 50 Schüler*innen eines deutschen Gymnasiums zeigt, dass besonders das tutororientierte Modell positive Rückmeldungen hinsichtlich Nutzerfreundlichkeit und wahrgenommenem Lernwert erhielt. Gleichzeitig bleibt Skepsis gegenüber der Genauigkeit und dem schulischen Einsatz solcher Systeme bestehen. [1]

Parallel dazu untersuchte Ogbo-Gebhardt et al. [16] den Einsatz eines LLM-Assistenzsystems mit RAG bei einer unterrepräsentierten Studierendengruppe an einer US-amerikanischen Historically Black Colleges and Universities (HBCU). Der KI-Assistent wurde über WhatsApp eingebunden und auf studienrelevante Inhalte begrenzt, um Halluzinationen zu vermeiden und pädagogische Kontrolle zu gewährleisten. Die qualitative Analyse zeigte, dass insbesondere Neugier, wahrgenommener Nutzen und einfache Zugänglichkeit entscheidend für die Akzeptanz waren. Dabei wurde auch deutlich, dass individuelle technische Vorkenntnisse und institutionelle Rahmenbedingungen maßgeblich das Nutzungsverhalten prägen. [16]

Ein weiteres Beispiel für die Anwendung dieser Technologie ist BARKPLUG V.2, ein Chatbot-System, das an der Mississippi State University entwickelt wurde. Das System nutzt universitätsspezifische Daten als externe Wissensquelle und verarbeitet diese mithilfe einer RAG-Pipeline. Die Systemarchitektur umfasst dabei zwei Hauptphasen, die kontextbasierte Dokumentenretrieval mittels Vektor-Datenbank und die darauffolgende Antwortgenerierung über ein GPT-basiertes Modell. Die Wirksamkeit von BARKPLUG V.2 wurde sowohl quantitativ mithilfe des RAGAS-Frameworks als auch durch eine Usability-Studie mittels SUS evaluiert, mit positiven Ergebnissen. [13]

In der Arbeit „An LLM-Driven Chatbot in Higher Education for Data Science-Related Topics“ von Neumann et al. [17] wird ein weiteres System vorgestellt, das einen domänenspezifischen Chatbot für datenwissenschaftliche Fragestellungen im akademischen Kontext entwickeln soll. Hierbei wurde ein LLM durch Retrieval-Komponenten erweitert, das Inhalte aus wissenschaftlichen Artikeln, universitären Ressourcen und Open-Access-Datenbanken wie ArXiv und PubMed abrufen konnte. Die Evaluation erfolgte sowohl durch eine technische Analyse der Modellantworten (z. B. Korrektheit und Relevanz) als auch durch Nutzerbefragungen. Die Ergebnisse zeigten, dass das System in der Lage war, qualitativ hochwertige Antworten zu generieren, die den Erwartungen der Zielgruppe, Studierende und Lehrende im Data-Science-Umfeld, weitgehend entsprachen. [17]

Im Projekt von Barkana et al. [20] wurde ein interaktives Problemlösungs-Szenario entwickelt, das die Zusammenarbeit zwischen fachfremde Nutzer*innen und einem humanoiden NAO-Roboter untersucht. Zentral war die Integration des erklärbaren Dialog-Managers DAISY, der Nutzenden sowohl rationale als auch leicht nachvollziehbare Erklärungen während der Interaktion bietet. Die Studie verglich zwei Bedingungen, einmal den NAO-Roboter mit DAISY und zum anderen eine rein webbasierte DAISY-Version ohne physische Robotik. In beiden Fällen wurde das gleiche Aufgabenformat verwendet, bei dem die Teilnehmenden mit dem System zusammen logische Probleme lösen mussten. Ziel war es, den Einfluss der physisch-sozialen Präsenz des Roboters auf die Benutzererfahrung und -präferenz zu evaluieren. Insgesamt nahmen 25 Versuchspersonen an der Studie teil, wobei 18 von ihnen am Ende die

NAO-gestützte Variante bevorzugten. Die SUS-Werte unterstützten diese subjektive Präferenz, da der NAO mit einem Mittelwert von 73,1 gegenüber 68,8 beim Websystem besser abschnitt. Besonders hervorgehoben wurde die natürlichere Interaktion durch Sprache und Körpersprache sowie die verbesserte Aufmerksamkeit der Teilnehmenden. Methodisch wurde sowohl quantitativ über Fragebögen als auch qualitativ über Beobachtungen und Interviews ausgewertet. [20]

In der Studie "The effect of a physical robot on vocabulary learning" von Wedenborn et al. [21] wurde untersucht, wie sich die physische Verkörperung eines Tutors auf den Spracherwerb auswirkt. Dafür entwickelten die Forscher eine Anwendung zur Vermittlung russischer Vokabeln in drei Bedingungen, mit einem körperlosen Sprecher, einem animierten Avatar und einem physischen Roboterkopf (Furhat). Die 15 Studierende ohne Vorkenntnisse lernten je neun Wörter in jeder Bedingung. In einem anschließenden Test sollten sie die gehörten Wörter den passenden Bildern zuordnen. Die Studie verwendete eine Wizard-of-Oz-Methode zur Bewertung der Aussprache, wobei ein versteckter Mensch das korrekte Nachsprechen bestätigte. Die Ergebnisse zeigten signifikant bessere Lernergebnisse in der Roboterbedingung verglichen mit den anderen. Die Forschenden führen dies auf eine höhere Motivation, stärkere emotionale Bindung und die Neuartigkeit des physischen Roboters zurück. Ein Fragebogen bestätigte zudem, dass Teilnehmende die Roboterinteraktion als unterhaltsamer und persönlicher wahrnahmen. [21]

Im Rahmen der Studie „From Robots to Chatbots: Unveiling the Dynamics of Human-AI Interaction“ untersuchten Lukasik et al. [22] die unterschiedlichen Wirkungen physisch verkörperter Agenten, wie humanoider Roboter, im Vergleich zu rein textbasierten Chatbots. Die Ergebnisse zeigen, dass humanoide Roboter in Interaktionen als emotionaler und sozial präsenter wahrgenommen werden. Diese Form der sozialen Präsenz kann positive Effekte auf Motivation und Engagement haben. Gleichzeitig wird jedoch deutlich, dass diese emotionalen Vorteile nicht zwingend mit einer besseren Bedienbarkeit oder Effizienz einhergehen. Textbasierte Chatbots werden hingegen routinierter und kognitiv zielgerichteter genutzt, was sich vor allem in strukturierten Anwendungsszenarien zeigt. Die Studie betont damit die komplementären Stärken beider Systeme, während Roboter in sozialen Lern- oder Betreuungskontexten Vorteile bieten, überzeugen Chatbots durch pragmatische Einfachheit. Diese Ergebnisse unterstreichen die Relevanz der Kontextabhängigkeit bei der Auswahl von KI-Systemen im Bildungsbereich. In didaktischen Szenarien mit starkem Fokus auf zwischenmenschliche Interaktion könnte der Einsatz eines Roboters sinnvoller sein, während in selbstgesteuerten, wissensbasierten Settings der Web-Chatbot Vorteile bietet. Die Forschung liefert somit einen differenzierten Beitrag zur Diskussion um die Effektivität verkörperter versus digitaler KI-Agenten in der Bildung. [22]

Laut Zhou et al. [23] beeinflussen Leistungserwartung, einfache Bedienbarkeit, sozialer Einfluss und hedonische Motivation maßgeblich die Verhaltensabsicht zur Nutzung von Chatbots. Das erweiterte UTAUT2-Modell dient dabei als theoretische Grundlage zur Erklärung der Akzeptanz digitaler Assistenten. Zusätzlich berücksichtigt die Studie das Task-Technology-Fit-Modell, um die Passung zwischen den Funktionen des Chatbots und den Anforderungen der jeweiligen Aufgaben zu analysieren. Eine hohe Passung sowie ein klar strukturierter, nutzerorientierter Kommunikationsstil fördern das Vertrauen und die tatsächliche Nutzung. Technologiebezogene Ängste und Datenschutzbedenken wirken sich hingegen hemmend auf die Nutzungsbereitschaft aus. Insgesamt zeigt sich, dass die erfolgreiche

Implementierung von Chatbots nicht nur technische, sondern vor allem psychologische und soziale Faktoren erfordert. [23]

Tarlan et al. [24] untersuchten die Reaktionen von Nutzenden auf den humanoiden Roboter Nao im Vergleich zum virtuellen Avatar Jack. Dabei zeigte sich, dass der Roboter als sympathischer und geringfügig intelligenter wahrgenommen wurde. Allerdings ergaben die Ergebnisse keinen signifikanten Unterschied in der grundsätzlichen Akzeptanz oder im wahrgenommenen Anthropomorphismus beider Systeme. Das bedeutet, dass der humanoide Roboter trotz seines physischen Erscheinungsbildes gegenüber dem Webchatbot keinen klaren Vorteil hinsichtlich der Nutzerakzeptanz aufwies. Diese Befunde legen nahe, dass visuelle oder körperliche Merkmale allein nicht ausschlaggebend für die Akzeptanz sind. Vielmehr spielen funktionale Merkmale und das individuelle Nutzungserlebnis eine zentrale Rolle. [24]

Trotz der zunehmenden Verbreitung von KI-Technologien im Bildungsbereich existieren bislang kaum direkte Vergleichsstudien zwischen humanoiden Robotern wie NAO und webbasierten Chatbots mit RAG. Die existierende Literatur fokussiert sich häufig entweder auf humanoide Roboter in sozialen und lernunterstützenden Kontexten oder auf Chatbots mit generativer KI wie ChatGPT jedoch nicht auf systematische, empirisch fundierte Gegenüberstellungen beider Technologien im gleichen experimentellen Setting. Insbesondere der Einsatz von RAG-gestützten Web-Chatbots in Bildungskontexten wurde bislang primär isoliert untersucht, wodurch Aussagen über differenzielle Wirkmechanismen oder Akzeptanzmuster im direkten Vergleich mit Robotersystemen nur eingeschränkt möglich sind. Dieser Forschungsengpass unterstreicht die Relevanz und den innovativen Beitrag der vorliegenden Arbeit.

2.2 KI in der Bildung

Generative KI-Systeme (GKI) wie ChatGPT oder RAG-Pipelines eröffnen neue Möglichkeiten für den Bildungsbereich. Durch ihre Fähigkeit, natürliche Sprache zu erzeugen und interaktiv zu kommunizieren, können sie Lernprozesse bereichern und die Umsetzung des Sustainable Development Goals (SDG) 4 erhöhen, vorausgesetzt, ihr Einsatz erfolgt reflektiert und verantwortungsvoll. Im Bildungsbereich hat Technologie den Unterricht und das Lernen grundlegend verändert, dank virtueller Plattformen, digitaler Materialien und interaktiver Anwendungen. Diese technische Einbindung gilt als bedeutender Fortschritt, da sie das Lernerlebnis der Studierenden ebenso bereichert wie die Unterrichtspraxis der Lehrenden. [5]

Peyton et al. [2] heben hervor, dass Chatbots im Bildungsbereich vielfältige Vorteile bieten könnten, darunter personalisierte Unterstützung und ständige Verfügbarkeit rund um die Uhr. Dennoch konzentriert sich der Großteil bisheriger Forschung und praktischer Anwendungen von Chatbots vor allem auf Branchen wie Telekommunikation, Finanzwesen und E-Commerce. [2]

2.2.1 Wichtigkeit der KI in der Bildung

KI hat sich in den letzten Jahren zu einer Schlüsseltechnologie entwickelt, die das Potenzial besitzt, das Bildungswesen grundlegend zu transformieren. Eine solide Bildung legt das

Fundament für individuelles Wohlergehen und gesellschaftliche Entwicklung. Die Agenda 2030 der Vereinten Nationen hebt hervor, dass Bildungschancen verbessert und Ungleichheiten abgebaut werden müssen. Das vierte Nachhaltigkeitsziel fordert daher, bis 2030 eine inklusive, gerechte und hochwertige Bildung sicherzustellen und allen Menschen lebenslanges Lernen zu ermöglichen. [25]

Insbesondere die Integration von KI-Systemen in Lehr- und Lernprozesse ermöglicht personalisierte, adaptiv gesteuerte Lernpfade, welche auf die individuellen Bedürfnisse und Vorkenntnisse der Lernenden eingehen. Diese Form der Adaptive Learning-Technologie (ALT) steigert nachweislich die Lernergebnisse. Studien zeigen, dass KI-gestützte adaptive Tutorien die Testergebnisse von Studierenden um 62% verbessern können. [26]

Ein zentrales Argument für den Einsatz von KI im Bildungsbereich ist die Fähigkeit zur Skalierung von Tutoring und Feedback. Während traditionelle Lernumgebungen auf Lehrkräfte mit begrenzter Verfügbarkeit angewiesen sind, können KI-gestützte Systeme wie „AutoTutor“ oder der „Socratic Playground“ hunderttausende Student*innen oder Schüler*innen zeitgleich individuelle Unterstützung leisten und komplexe Rückmeldungen in Echtzeit geben. Dadurch reduziert sich nicht nur der Arbeitsaufwand der Lehrenden, sondern es entstehen auch leichter Zugänge zu qualitativ hochwertiger pädagogischer Betreuung. [27]

Zudem fördert KI die Daten- und Lernanalyse durch die systematische Auswertung großer Datenmengen, die in digitalen Lernumgebungen generiert werden. So können Lehrende frühzeitig Lernschwierigkeiten erkennen und gezielt intervenieren, was zu einer Senkung der Abbrecherquoten und einer höheren Abschlussrate von Bildungsprogrammen führt. Human-Centred Learning Analytics (HCLA)-Studien betonen, dass die Einbindung von Stakeholdern im Design dieser Systeme unerlässlich ist, um Vertrauen und Akzeptanz zu gewährleisten. [28]

Neben diesen Vorteilen trägt KI im Bildungssektor auch zur Förderung von Chancengleichheit bei, wie dem SDG 4. Personalisierte Lernumgebungen gleichen Bildungslücken aus, indem sie Lerninhalte adaptiv an das Tempo und die Vorkenntnisse einzelner Lernender anpassen. Insbesondere in strukturschwache Regionen können mobile KI-Applikationen Basisbildung ermöglichen, ohne dass es einer vollwertigen Infrastruktur bedarf. [29]

2.2.2 Grundlagen der Bildung im Kontext KI-gestützter Systeme

Laut einer systematischen Metaanalyse, die Artificial Intelligence in Education (AIEd)-Forschung im Hochschulbereich zusammenfasst, ist die rasche Entwicklung und Publikationsdynamik ein Indikator dafür, dass eine fundierte theoretische Verankerung unabdingbar ist [30]. Bereits in der Mathematik, Informatik, Naturwissenschaft und Technik (MINT)-Bildung wurde in einer systematischen Studie eine Struktur identifiziert, welche die Elemente eines KI-MINT-Systems entlang von Informationsfluss, Lerninhalten, Medium und Lernumgebung kategorisiert [31]. Diese systemtheoretische Perspektive bietet ein klares Framework, um KI-Komponenten im Bildungskontext zu verorten.

Die unterschiedlichen Ausprägungen KI-gestützter Anwendungen im Bildungsbereich lassen sich in drei Paradigmen einordnen, nämlich AI-directed, AI-supported und AI-empowered

Systeme. Diese Klassifizierung wurde in einem Artikel von Ouyang et al. [32] vorgestellt. AI-directed Systeme sind durch ein behavioristisches Lernverständnis geprägt. Hier übernimmt die KI die Rolle des „Lehrers“, steuert Lerninhalte und -prozesse und der Lernende ist hauptsächlich Empfänger vorgegebener Lernpfade. AI-supported Systeme beruhen stärker auf kognitivem und sozial-konstruktivistischem Lernen. Die KI unterstützt den Lernprozess, während der Lernende aktiv mit der KI zusammenarbeitet und durch Interaktion ein individuelleres, lernendenzentriertes Lernen ermöglicht wird. AI-empowered Systeme stellen die Lernenden ins Zentrum, die als aktive Gestalter*in ihres Lernens auftreten. Hier dient die KI als Werkzeug, um menschliche Intelligenz zu erweitern, personalisiertes Lernen zu fördern und eine hohe Eigenverantwortung der Lernenden zu ermöglichen. [32]

Ein weiterer grundlegender Aspekt ist die KI-Kompetenz, insbesondere vom Kindergarten bis zur 12. Klasse (K-12-Bereich). Bildungssysteme stehen vor der Aufgabe, Schüler*innen so auszubilden, dass sie selbstbestimmt und reflektiert mit KI-Systemen interagieren können [33]. Diese Kompetenzentwicklung stellt eine bildungspolitische und kognitive Herausforderung dar.

Eine Studie der Hochschullehre zeigen, dass Anwendungen wie intelligente Tutoren, automatisierte Bewertungssysteme, personalisierte Lernpfade und adaptive Lernumgebungen zentrale Bildungsinnovationen darstellen [34].

Eine historische Perspektive betont, dass die frühesten KI-gestützten Bildungssysteme auf Konzepten wie dem Programmed Logic for Automatic Teaching Operations (PLATO)-System in den 1960ern aufbauen und in den 1980ern und 1990ern zu den intelligente tutorielle Systeme (ITS) führten [35]. Diese Entwicklung unterstreicht, dass heutige KI-gestützte Bildung auf einem jahrelangen Paradigmenwechsel basiert, der von Rechenunterstützung hin zu lernzentrierter Personalisierung führt.

Die Förderung von Gerechtigkeit und Chancengleichheit durch AIEd ist ein häufig genanntes Ziel. So eröffnen KI-gestützte Systeme das Potenzial, individuelle Förderung im großen Maßstab zu ermöglichen und bisher isolierte Lerngruppen besser zu erreichen. Gleichzeitig mahnt Holstein et al. [36], dass bestehende Systemungleichheiten ohne sorgfältige Gestaltung auch verstärkt werden könnten.

2.2.3 Vergleich traditioneller vs. KI-gestützter Bildung

Der Vergleich zwischen traditioneller und KI-gestützter Bildung zeigt fundamentale Unterschiede in Pädagogik, Lernzugang und Effizienz. Eine Studie im Bereich der Programmierung belegt, dass AI-unterstütztes Pair Programming (PP), z. B. mit GPT-3.5 oder Claude 3 Opus, im Vergleich zu menschlichem PP die intrinsische Motivation deutlich steigert, Angst reduziert und die Programmierleistung verbessert, wobei die Ergebnisse den traditionellen Methoden überlegen waren [37]. In adaptiven Lernsystemen wie "Yixue Squirrel AI" erreichten Schüler*innen in Mathematik und Englisch bessere Leistungen als in traditionellem Frontalunterricht durch erfahrene Lehrkräfte [38]. Diese KI-Systeme ermöglichen eine personalisierte, datengetriebene Lernsteuerung, die auf den aktuellen Wissensstand jedes Lernenden zugeschnitten ist.

Im Gegensatz dazu liegt der Schwerpunkt traditioneller Bildung häufig auf standardisierten Lehrplänen und persönlicher Interaktion, was zwar soziale Kompetenzen fördert, aber

individuelle Anpassung erschwert [39]. Laut einer Studie von Henze et al. [40] unterstützte KI, Studierende bei datenintensiven Aufgaben wie Pendel-Experimenten, wobei Lernzuwachs ähnlich, aber Motivation deutlich höher war als bei traditionellen Excel-basierten Methoden.

Vergleichende Studien im Bereich der Bildungsmedien zeigen, dass KI-generierte und menschlich erstellte Lehrvideos ähnliche Lernergebnisse erzielen [41]. Menschliche Videos werden jedoch subjektiv leicht bevorzugt, insbesondere aufgrund ihrer stärkeren sozialen Präsenz [41]. Eine Studie von Buthoria et al. [42] zeigt, dass KI vor allem in der Personalisierung und adaptiven Ausgestaltung von Lernpfaden stärker ausgeprägt sind, allerdings ist eine begleitende menschliche Steuerung entscheidend, um Qualität und Relevanz sicherzustellen.

Insgesamt zeigt sich ein differenziertes Bild, da KI-gestützte Bildung Effizienz, Motivation und personalisierte Unterstützung verbessert, während traditionelle Methoden soziale Interaktion, kritisches Denken und persönliche Bindung fördern. So entsteht ein integrativer Bildungsansatz, der Skalierbarkeit, Individualität und Tiefenlernen zugleich ermöglicht, idealerweise evidenzbasiert und motiviert durch didaktische Klarheit.

2.2.4 Verbindung von KI und Bildung

Systematische Delphi-Analysen zeigen, dass AIED inzwischen als Schlüsseltechnologie für die Transformation formaler, non-formaler und informeller Lernsettings gilt und dabei insbesondere Fragen der Human-in-the-Loop-Kooperation (HITL) in den Vordergrund rücken [43].

Andererseits fungiert KI als Werkzeug zur Unterstützung von Lehr-Lern-Prozessen, etwa mittels adaptiver Systeme, automatisierter Bewertung oder intelligenter Tutoring-Systeme [44]. Adaptive Lernplattformen machen Lernwege stärker personalisiert, steigern Schüler*innen-Engagement und verbessern Leistung, dies gilt als exemplarischer Einsatz von KI im e-Learning [45].

Ethikthemen wie Fairness, Verantwortung und Datenintegrität sind eng mit der Gestaltung von KI-Bildungssystemen verknüpft, eine Implementierung erfordert explizite ethische Reflexion [33]. Anstatt ausschließlich auf vortrainierte Modelle zu vertrauen, kombiniert RAG externe, geprüfte Wissensquellen mit generativer Verarbeitung, wodurch Fairness und Verantwortlichkeit gestärkt werden. Gleichzeitig reduziert sich das Risiko von halluzinierten Antworten und fehlerhafter Informationsvermittlung, was gerade in sensiblen Bildungsumgebungen fundamental ist. Damit leisten RAG-Pipelines einen Beitrag zur ethisch reflektierten Implementierung von KI im Bildungsbereich. [46]

Eine mögliche Verbindung von KI und Bildung zeigt sich im Einsatz von Chatbots und humanoiden Robotern, die kognitive, soziale und emotionale Dimensionen des Lernens zugleich adressieren.

Humanoide Roboter wie der NAO-Roboter bringen eine physische, soziale und emotionale Komponente in Lernsettings, daraus resultiert das man diese Technologie in der Bildung mit KI verbinden sollte [47]. In einem explorativen Einsatz in der Hochschulbildung, wurde der NAO-Roboter als freundlich, sympathisch, motivierend und vertrauenswürdig wahrgenommen, und Studierende schätzten ihn als positiven Beitrag zum Engagement und Lernerfolg [48]. Ein

besonders innovatives Feld ist der hybride Einsatz, beispielsweise vernetzte Lernräume mit Mensch-KI-Kollaboration. In Peking wurde ein intelligenter Klassenraum mit einem Bildungsroboter eingesetzt, um kollaborative Lernsettings zu unterstützen [49].

Neueste Studien belegen, dass LLMs in Verbindung mit domänenspezifischen Wissensgraphen lernerspezifische Erklärungen erzeugen können, ohne auf starre Item-Pools zurückzugreifen. Bei Lese- und Mathematik-Lernständen verbesserte sich die Passung zwischen Lernziel und Material um durchschnittlich 27 %. [50]

2.3 Web-Chatbot

Ein Web-Chatbot ist ein softwarebasiertes System, das über eine Webschnittstelle menschenähnliche Konversationen mittels natürlicher Sprache ermöglicht. Laut Chitto et al. [51] und Sidlauskiene et al. [52] ist ein Chatbot als "software-based system designed to interact with humans using natural language" definiert. Typischerweise kommt dabei Natural Language Processing (NLP) zum Einsatz, wodurch Eingaben des Nutzers verstanden und angemessen verarbeitet werden können. Laut Adamopoulou et al. [53] ist ein Webchatbot, ein Computersystem, das mittels KI mit Anwender*innen in natürlicher Sprache interagiert und dabei den Eindruck eines intelligenten Gesprächspartners vermittelt. Ein Web-Chatbot fungiert als Vermittler zwischen Nutzer*innen und Website, übernimmt Aufgaben in der grafischen Benutzeroberfläche und informiert über den Stand der Interaktion [51].

2.3.1 Web-Chatbot in der Bildung

Webbasierte Chatbots haben sich als vielseitiges Werkzeug im Bildungsbereich etabliert und bieten durch ihre Plattformunabhängigkeit und jederzeitige Verfügbarkeit neue Möglichkeiten zur Unterstützung von Lern- und Lehrprozessen. Die Studie von Cai et al. [54] verdeutlicht, dass webbasierte Chatbots nicht nur für einfache FAQ-Funktionen eingesetzt werden, sondern zunehmend, als kollaborative Lernassistenten fungieren, die Gruppenarbeiten moderieren, Diskussionsanreize setzen und peer-to-peer-Austausch fördern können. Studien von Kurni et al. [55] und Ait et al. [56] heben hervor, dass die Integration von KI in Web-Chatbots das Lernengagement signifikant steigern können, da verschiedene Lerntypen und sensorische Kanäle angesprochen werden.

Darüber hinaus zeigen Untersuchungen, dass Web-Chatbots Lernprozesse nicht nur durch Informationsbereitstellung, sondern auch durch adaptive Unterstützung individualisieren können, indem sie den Wissensstand der Lernenden berücksichtigen [57]. Web-Chatbots können hier als Gesprächspartner agieren, der unbegrenzt Übungsmöglichkeiten bietet und Fehler unmittelbar korrigiert. Zudem tragen sie durch Gamification-Elemente wie Punkte, Abzeichen oder Fortschrittsanzeigen zur Steigerung der Motivation bei [58].

In inklusiven Bildungsumgebungen ermöglichen Chatbots den barrierearmen Zugang zu Lernressourcen. Dies fördert die Teilhabe von Personen mit unterschiedlichen Bedürfnissen und Lernvoraussetzungen. Gleichzeitig können Lehrende über die Interaktionsprotokolle wertvolle Einblicke in Lernschwierigkeiten und Wissenslücken erhalten, was eine gezieltere Unterrichtsplanung ermöglicht. [59]

Die Kombination aus Chatbots und RAG steigert den Nutzen enorm, da die Systeme damit in der Lage sind, aktuelle und fachlich verlässliche Informationen aus externen Datenquellen einzubinden. Dadurch wird vermieden, dass Antworten auf veralteten oder unvollständigen Trainingsdaten basieren, was insbesondere in dynamischen Domänen wie Informatik oder Medizin entscheidend ist. RAG stärkt somit die Genauigkeit und Aktualität generativer Antworten. Diese Methode ist besonders relevant für Bereiche, in denen fachliches Wissen rasch wächst und sich kontinuierlich wandelt. [60]

Abschließend lässt sich festhalten, dass Web-Chatbots im Bildungsbereich weit mehr als ein reines Informationswerkzeug darstellen, sie sind integraler Bestandteil moderner Lernökosysteme, die sowohl Individualisierung, Interaktivität als auch skalierbare Unterstützung bieten. Ihre pädagogische Wirksamkeit hängt jedoch entscheidend von einer didaktisch fundierten Implementierung, technischer Qualität und der kontinuierlichen Anpassung an die Bedürfnisse der Lernenden ab.

2.3.2 Bisherige Studien und Anwendungen von Web-Chatbots in Schulen und Hochschulen

Web-Chatbots haben sich in den letzten Jahren als vielseitige Werkzeuge zur Unterstützung von Lehr-Lern-Prozessen etabliert. Erste systematische Untersuchungen in Schulen wiesen bereits auf das Potenzial hin, die Schüler-Motivation und das Engagement sich dadurch steigert. So zeigte eine Studie von Benotti et al. [61] in zwei argentinischen Hochschul-Kursen, dass Chatbot-gestützter Unterricht im Vergleich zu traditionellen Umgebungen höhere Rückhalte- und Interessensraten, insbesondere bei weiblichen Schüler*innen, erzielte. Ein weiterer Ansatz wurde im Rahmen des „Smart School Framework“ präsentiert, bei dem ein Web-Chatbot in einem Internet of Things (IoT)-gestützten Schulumfeld Umweltdaten zugänglich gemacht und so sowohl naturwissenschaftliches Interesse als auch das Informatik-Verständnis förderte [62].

Merelo et al. [63] analysierten die Lehrerperspektiven auf den Einsatz von Chatbots und Messaging-Plattformen in der Hochschullehre. Die Befragung von Lehrenden in spanischsprachigen Ländern identifizierte als Hauptvorteile die Möglichkeit einer orts- und zeitunabhängigen Lernbegleitung, aber auch Herausforderungen wie erhöhten Schulungsaufwand für Dozent*innen und Datenschutzfragen. Eine umfassende Studie von Taylor et al. analysierte die historische Entwicklung der Chatbot-Integration auf US-Hochschulwebsites von 2017 bis 2023 und stellte heraus, dass Informations- und Beratungsanwendungen häufigsten Einsatzszenarien bildeten [64].

In vergleichbaren Szenarien setzten Forschende an der University of Michigan „ARGObot“ ein, der Studierenden automatisierte Studienberatung anbot, indem er universitätsspezifische Richtlinien interpretierte. Die Evaluation ergab, dass ARGObot die Navigation komplexer Regularien erleichterte und die Zufriedenheit der Nutzer deutlich erhöhte [65].

Hinzu kommen spezialisierte Chatbots für psychologische und sozio-emotionale Unterstützung. Jusoh et al. [66] entwickelten „HelpBot“ zur Bewältigung von Depressionen bei Jugendlichen. In schulischen Pilotversuchen reduzierte der Bot nachgewiesenermaßen das Stressempfinden und förderte die Bereitschaft, professionelle Hilfe in Anspruch zu nehmen.

Dieses Beispiel verdeutlicht, dass Chatbots nicht nur kognitive Prozesse, sondern auch das Wohlbefinden von Lernenden positiv beeinflussen können. [66]

2.3.3 Technologische Grundlagen von Web-Chatbots

Kernstück moderner Chatbots sind Natural-Language-Processing-(NLP-)Technologien für Sprachverstehen und Sprachgenerierung. Zunächst analysiert ein Natural Language Understanding (NLU) Modul die Chatbot-Anfrage, um die Benutzerabsicht (Intent) und relevante Entitäten (Schlüsselwörter, Parameter) zu erkennen. Auf Basis dieser Interpretation entscheidet die Dialogverwaltung, wie weiter vorgegangen werden soll, ob eine Datenbankabfrage nötig ist, zusätzlicher Kontext abgewartet wird oder eine Rückfrage zur Klärung gestellt werden muss. Gegebenenfalls ruft der Chatbot externe Informationen ab, zum Beispiel aus einer Wissensdatenbank oder über ein Web-Application-Programming-Interface (API), um die Anfrage zu beantworten. Anschließend formuliert ein Natural Language Generation (NLG) Modul die Antwort in natürlicher Sprache, die dem Anwender*innen über das Frontend präsentiert wird. Die Dialog-Management-Komponente hält dabei den Kontext des Gesprächs (erkannter Intent, zuvor genannte Entitäten etc.) fest, um konsistente mehrzügige Dialoge zu ermöglichen und Folgeverhalten zu steuern. [53]

Moderne Web-Chatbots nutzen immer häufiger LLM's, große vortrainierte Sprachmodelle wie z. B. GPT-3/4/5, zur Generierung von Antworten. Solche transformer-basierten Modelle ermöglichen qualitativ hochwertige und kontextbezogene Dialogbeiträge und können auch komplexe Anfragen bewältigen. Allerdings besteht bei reinen LLM-basierten Chatbots die Gefahr von Halluzinationen, also Antworten, die zwar sprachlich überzeugend klingen, inhaltlich aber falsch oder unbegründet sind. Um dies zu verhindern, setzen aktuelle Systeme auf RAG. Dieser Ansatz kombiniert ein generatives Sprachmodell mit einer nachgeschalteten Informationssuche, sodass vor der Antwortgenerierung relevantes externes Wissen aus Datenbanken oder dem Web abgerufen und in die Antwort eingebettet wird. Durch diese Verknüpfung von LLM und Wissensdatenbank kann die Faktentreue der Antworten deutlich verbessert und das Risiko von Halluzinationen verringert werden. [67]

Die Umsetzung von Web-Chatbots wird durch zahlreiche Plattformen und Frameworks unterstützt, die eine Anbindung an Webanwendungen erleichtern. So bieten große Cloud-Anbieter konversationsorientierte NLP-Dienste (etwa Dialogflow, IBM Watson Assistant oder Microsoft LUIS) an, die über Web-APIs in eigene Anwendungen integriert werden können. Daneben existieren Open-Source-Frameworks wie Rasa, die Entwicklern erlauben, eigene Dialogsysteme zu erstellen und diese z. B. via REST-Schnittstellen oder Software Development Kit (SDK) in Webseiten und Messenger zu integrieren.

2.4 Humanoide Roboter

Als humanoide Roboter bezeichnet man Maschinen in bipedaler, menschenähnlicher Form, die dafür konzipiert sind, in menschbezogenen Umgebungen zu arbeiten und mit Menschen zu interagieren [68], [69]. Humanoide Roboter zeichnen sich durch eine verkörperte soziale Präsenz aus, eine Eigenschaft, von der man annimmt, dass sie sie zu idealen Plattformen für den Einsatz im Bildungsbereich macht [70]. Studien im Kontext der HRI zeigen, dass durch

die menschenähnliche Gestaltung (Gestik, Mimik, Körperhaltung) und das gezielte Design sozialer Interaktion Lernende stärker einbezogen werden [70].

In anatomischen Trainingsszenarien können humanoide Roboter komplexe Demonstrationen von Bewegungsabläufen zeigen und so das Verständnis für menschliche Anatomie deutlich verbessern [71].

Durch das Mirroring sozialer Signale können sie Empathie aufbauen. Mirroring (Spiegeln) bezeichnet in der Sozial- und HRI-Forschung das meist unbewusste Nachahmen der nonverbalen Signale eines Gegenübers. In der MRI umfasst es affektives Mirroring (Gefühlsausdrücke/LED-Signale) und Bewegungs-Mirroring (Kopf/Blick) und wird u. a. mit Mechanismen des Spiegelneuronensystems begründet, dadurch wirkt der Roboter sozial näher, aufmerksamer und menschlicher. Empirische Befunde zeigen, dass solches Spiegeln die wahrgenommene Empathie, Responsivität und Menschlichkeit des Roboters steigern kann. [72]

Die menschenähnliche Gestaltung stellt zugleich eine Herausforderung dar, da ein zu realistisch wirkendes Erscheinungsbild bei gleichzeitig subtil „nicht menschlich“ wahrgenommenem Verhalten oder Ausdruck Abwehrreaktionen hervorrufen kann, ein Effekt, der als Uncanny Valley bekannt ist.

2.4.1 Uncanny Valley

Der Begriff Uncanny Valley wurde 1970 vom Robotiker Masahiro Mori geprägt und bezeichnet ein Phänomen in Robotik und Animation. Menschen reagieren dabei zunächst positiv auf menschenähnliche Darstellungen, empfinden aber plötzlich Abneigung oder Unbehagen, sobald diese fast, aber nicht ganz menschlich wirken. [73]

Dieses Phänomen tritt insbesondere bei Robotern, Avataren, digitalen Figuren oder anderen menschenähnlichen Entitäten auf. Die Diskrepanz zwischen dem fast perfekten menschlichen Äußeren und subtilen, unnatürlich wirkenden Bewegungen oder Gesichtsausdrücken führt bei den Anwender*innen zu Verunsicherung und Irritation. Studien haben gezeigt, dass schon kleine Unstimmigkeiten in der Mimik oder im Bewegungsablauf ausreichend sind, um eine Gestalt als „unheimlich“ wahrzunehmen und zu meiden. Durch dieses Szenario sinken Sympathie und Vertrauen gegenüber dem Robotikssystem signifikant. [74]

Kim et al. [75] zeigen anhand von 251 Ganzkörperbildern realer Roboter, dass die Bewertungsverläufe über den Human-Likeness-Score am besten durch ein quartisches Modell erklärt werden, mit zwei Wendepunkten, also zwei Uncanny Valleys. Das kleinere Tal liegt bei moderat niedriger Menschenähnlichkeit, wenn wenige Gesichts- und Oberflächenmerkmale auf stark ausgeprägte Körper-/Manipulator-Merkmale treffen (Low Surface/Face gegenüber einem High Body-Manipulator). Das klassische Mori-Tal erscheint dagegen bei hoher, aber nicht perfekter Menschenähnlichkeit, wenn viele Gesichts-/Oberflächenmerkmale mit einem vergleichsweise wenig menschenähnlichen Körper kollidieren (High Surface/Face gegenüber einem Low Body-Manipulator). Zudem sind beide Täler in negativen Reaktionen deutlich ausgeprägter als in positiven, was die Bedeutung solcher Wahrnehmungsdiskrepanzen gegenüber bloßer Gesamt-Menschenähnlichkeit unterstreicht. [75]

Neben klassischen Erklärungsansätzen werden zunehmend auch kognitive Mechanismen diskutiert, die das Uncanny-Valley-Phänomen erklären sollen. Moore [76] zeigt in einem bayesianischen Modell, dass Irritationen im Uncanny Valley durch perzeptuelle Konflikte entsteht, wenn Wahrnehmungshinweise einer Figur, menschliche als auch nicht-menschliche Eigenschaften signalisieren, kommt es zu einem psychischen Spannungszustand. Diese Inkonsistenz löst beim Betrachter Gefühle von Unsicherheit oder Abwehr aus, die als zentrale Ursache des Uncanny-Valley-Effekts gelten können [76].

Kätsyriet et al. [77] argumentieren, dass das Phänomen auf Kategorisierungsprobleme im Gehirn zurückgeht, wenn eine Figur weder eindeutig als menschlich noch eindeutig als künstlich eingeordnet werden kann, führt dies zu kognitiver Dissonanz und Unbehagen.

2.4.2 Technologische Grundlagen des NAO-Roboters

Der NAO-Roboter, entwickelt von Aldebaran Robotics (heute SoftBank Robotics), vereint Mechatronik, ein verteiltes Elektronik- und Software-Ökosystem sowie ein Vollspektrum an Sensor- und Aktuatorsystemen zu einer kompakten humanoiden Plattform. Mit einer Körpergröße von 57 cm und einem Gewicht von rund 4,5 kg basiert der Mechatronik-Entwurf auf 25 Freiheitsgraden (Degrees of Freedom (DOF)), verteilt auf 14 DOF im Oberkörper (Arme und Kopf) und 11 DOF im Unterkörper (Beine und Becken) [78]. Die Gelenke werden über hocheffiziente Motoren angetrieben. Durch die Aggregation zweier Motor-Getriebe-Module zu einem universellen Gelenkmodul konnte die Zahl der benötigten Aktuatoren reduziert und der mechanische Aufbau vereinfacht werden, ohne Einbußen bei Drehmoment oder Präzision hinzunehmen [78].

Das Kopf-Mainboard verwendet einen x86-basierten AMD Geode-500 MHz-Prozessor mit 256 MB SDRAM und 1 GB Flash-Speicher für die echtzeitnahe Steuerung der Aktuatoren ist ein ARM7-60 MHz-Mikrocontroller zuständig, der über zwei RS-485-Feldbusse (jeweils 460 kbit/s) mit den Mikrocontroller-Modulen der Aktuatoren kommuniziert. Die RS-485-Segregation in Ober- und Unterkörper erhöht die Datendurchsatzrate und reduziert Latenzen bei Bewegungskontrollen. Die Verknüpfung zwischen CPU und ARM7 erfolgt über USB-2 mit bis zu 11 Mbit/s, wodurch Sensordaten aus externen Modulen in die Stabilitäts-Regelung einfließen können. [79]

Der Roboter NAO nimmt seine Umgebung mit einer Vielzahl von Sensoren wahr. Dazu gehören zwei Kameras für Stereo-Bilder, Ultraschallsensoren zur Abstandsmessung von Hindernissen sowie ein IMU-Modul mit Gyroskop und Beschleunigungssensor, das ständig die Körperlage erfasst. Drucksensoren in den Fußsohlen detektieren Bodenkontakt sowie Sturzereignisse und lösen automatisierte Aufsteh-Manöver aus. Ergänzt wird dies durch „Bumper“-Sensoren an Armen und Brust, die physische Interaktionen registrieren. [79]

Die Lauf- und Trajektoriensteuerung des Bipeds basiert auf einem hybriden Ansatz, der klassische Mustergeneratoren mit dynamischer Kompensationsregelung verbindet. Ein spezieller algorithmischer Kompensator passt die Stand- und Schwungphasen kontinuierlich an Störgrößen an und garantiert eine dynamisch stabile Fortbewegung selbst auf unebenem Untergrund. Real-Time-Experimente belegen, dass durch diesen Ansatz sowohl Schrittgeschwindigkeit als auch Robustheit gegenüber externen Kräften signifikant gesteigert werden können. [80]

Softwareseitig läuft NAO unter einem angepassten Linux-Kernel mit der NAOqi-Middleware, die eine einheitliche Schnittstelle für Applikationsentwicklung in C++, Python oder Java bietet. Die modular aufgebaute Architektur trennt Hoch- und Niedrig-Ebenen-Funktionalität, wobei zeitkritische Regelkreise für Gelenke und Sensoren auf den Mikrocontrollern laufen, während die Bild- und Sprachverarbeitung auf dem Hauptprozessor ausgeführt wird und über WLAN oder Ethernet Teleoperation, Debugging sowie Datenaustausch mit externen Systemen ermöglicht werden. [78]

Durch diese Kombination aus kompakten, leistungsfähigen Aktuatoren, vielseitigen Sensoren und einer verteilten Rechenarchitektur bietet NAO eine robuste, einfach programmierbare Plattform, die in Forschung, Lehre und Robotik eingesetzt wird. Ihre Offenheit und Modularität haben NAO zu einer der am weitesten verbreiteten humanoiden Roboterplattformen gemacht.

2.4.3 NAO im Bildungskontext

In der universitären Lehre findet NAO zunehmend Anwendung, beispielsweise in einem numerischen Methoden-Kurs, wo er eingesetzt wurde, um Studierende zu motivieren, sich aktiv mit Themen wie Vorwärts- und Inverskinematik auseinanderzusetzen. Innerhalb eines designbasierten Forschungsansatzes half der Roboter, die hypothetische Lerntrasse zur Vermittlung des Newton-Raphson-Verfahrens zu validieren und unterstützte reflexives Lehrerverhalten. [81]

Ein weiteres Einsatzfeld des NAO-Roboters zeigt sich in der Hochschullehre. In einem zweisemestrigen Informatikkurs für angehende Lehrkräfte diente NAO als praktisches Übungsmedium. Die Teilnehmenden schätzten besonders die intuitive Programmierumgebung Choregraphe und die Möglichkeit, durch reale Interaktion mit NAO Programmierkonzepte zu veranschaulichen. [82]

Eine Studie von Lyk et al. [83] untersucht NAO als Autoritätsperson im Unterricht. Hier zeigte sich, dass NAO durch seine Präsenz und klare akustische Ansagen dazu beitragen kann, die Lautstärke im Klassenzimmer zu kontrollieren und damit die Lernatmosphäre zu stabilisieren.

In der Sprachlehre fungierte NAO als Vokabeltrainer. In einem experimentellen Setting wurde gezeigt, dass Studierende den Roboter positiv bewerteten. Viele empfanden ihn als motivierend, freundlich und unterstützend beim Vokabellernen, auch wenn die statistisch signifikanten Lernerfolge begrenzt ausfielen. Der Roboter wurde vielfach menschlich wahrgenommen, etwa mit „er“ oder „mein Freund“ angesprochen und bot eine alternative, interaktive Ergänzung zum traditionellen Unterricht. [84]

Auch in der Vorschule kommt NAO zum Einsatz. Eine Studie von Perla et al. [85] beobachtete Kindergartenkinder während von Pädagog*innen vermittelten Aktivitäten mit NAO. Beide Gruppen, Erziehende und Eltern, berichteten über signifikante Verbesserungen in motorischen Fähigkeiten und gesteigerter Engagement-Motivation der Kinder. Darüber hinaus zeigte eine Nutzung in inklusiven Settings, insbesondere bei Schüler*innen mit Autismus-Spektrum-Störungen (ASS), vielversprechende Effekte, wie Aufmerksamkeit, Konzentration und Interaktionsverhalten verbesserten sich deutlich im Vergleich zu herkömmlichem Unterricht [86].

In der Hochschule kann NAO auch als „Scrum-Master“ fungieren, etwa in hybriden Lehr-Szenarien für agile Projektmanagementmethoden. In einer szenariobasierten Studie unterstützte er Daily-Scrum-Meetings. Studierende nahmen ihn als humanoides, anthropomorphes Wesen wahr. Dies förderte das die Beziehung zwischen Lernenden und Roboter und wurde als förderlich für den Lernprozess bewertet. Der multimodale Dialog mit Gestik, Stimme, Blicken und Bewegung trug entscheidend zur positiven Wahrnehmung bei. [87]

Zusammenfassend zeigt sich, dass NAO im Bildungskontext vielseitige Rollen übernehmen kann, wie zum Beispiel als motivierende Lernhilfe in technisch-wissenschaftlichen Lehrszenarien, als sprachfördernder, interaktiver Partner im Fremdsprachenunterricht, als Unterstützer in inklusiven Settings und Kindergarten, sowie als agiler Lernmoderator in Hochschulprojekten. Seine besondere Stärke liegt in seiner physischen Präsenz, multimodalen Ausdrucksfähigkeit und positiven humanoiden Wahrnehmung.

2.4.4 Potenziale und Einschränkungen humanoider Roboter im Unterricht

Humanoide Roboter bieten im modernen Unterricht bemerkenswerte Potenziale, da sie komplexe Sachverhalte anschaulich visualisieren und interaktive Lernumgebungen schaffen können. So können sie Schüler*innen oder Sudent*innen motivieren, in naturwissenschaftlich-technischen Fächern tiefer einzutauchen, indem sie Demonstrationen lebendig gestalten und kooperative Lerninteraktionen stimulieren [88]. Außerdem dienen sie als Hilfsmittel zur kognitiven Entlastung und können Schüler*innen bei der Bewältigung kognitiver Konflikte unterstützen, indem sie aktiv alternative Lösungswege aufzeigen oder visuelle Unterstützung bieten [69].

In der Praxis zeigten humanoide Roboter im Unterricht, zum Beispiel in zehn Schulen über drei Jahre hinweg, dass sie das Engagement von Lernenden steigern können, insbesondere wenn sie sinnvoll in pädagogische Abläufe integriert sind. Gleichzeitig wurde deutlich, dass ihr Einsatz auch logistische und didaktische Herausforderungen mit sich bringt, etwa in der Anpassung der Lehrpläne oder der Lehrkraft-Vorbereitung [47]. Auch in Sprachunterrichtsszenarien erwiesen sich humanoide Roboter als funktional und affektiv ansprechende Lernpartner, sie unterstützten sowohl kognitive als auch emotionale Komponenten des Lernens bei Grundschulkindern in Hongkong [14].

Parallel dazu stehen Einschränkungen des Einsatzes humanoider Roboter. So kann der Einsatz von Robotern in MINT-Unterricht insbesondere dazu führen, dass Schüler*innen durch technische Aspekte abgelenkt werden oder der kognitive Anspruch steigt, beides kann den eigentlichen Lernfortschritt mindern [89]. Darüber hinaus können Interaktionen mit Robotern soziale Normen verletzen und Verwirrung auslösen, wenn z. B. unpassende Gesten oder Verhaltensweisen auftreten, wie eine Studie zu kind-roboter Dialogen im Mathematikkontext zeigte [90].

Weitere Limitationen betreffen technische und gestalterische Aspekte. Die Entwicklung eines motivierend wirkenden Gestenrepertoires erweist sich als komplex, da bei Robotern Differenzen zwischen pädagogischer Absicht und realistischer Interaktion sorgfältig ausbalanciert werden müssen. [91]

Ein weiteres Hindernis besteht in der pädagogischen und organisatorischen Bereitschaft. Eine Studie, die das UTAUT-2- und Technology-Organization-Environment (TOE)-Modell kombiniert, macht deutlich, dass individuelle Faktoren (z. B. wahrgenommener Nutzen, Benutzerfreundlichkeit), organisatorische Unterstützung (z. B. Infrastruktur, Training) und Umweltbedingungen (z. B. politische Förderung) entscheidend für die tatsächliche Adoption humanoider Roboter durch Lehrkräfte sind. [92]

2.5 Retrieval-Augmented-Generation

Die Retrieval-Augmented-Generation-Methode (RAG) verfolgt das Ziel, Sprachmodelle durch das Abrufen und Einbinden relevanter Informationen aus externen Wissensdatenquellen zu optimieren und erzielt dabei spürbare Leistungssteigerungen bei wissensintensiven Aufgaben [93].

Die RAG fragt gezielt relevante Informationen aus einer Wissensdatenbank ab und bindet sie in den Eingabeprozess eines Sprachmodells ein. So kann das Modell Antworten auf Grundlage externer, evidenzbasierter Quellen erzeugen. Gleichzeitig wird damit häufig die Herausforderung entschärft, dass große Sprachmodelle als „Black Boxes“ agieren, deren interne Mechanismen weder transparent noch einfach nachvollziehbar sind. [94]

Technologisch nutzt RAG hochentwickelte Verfahren des NLP, um die Treffsicherheit und Aussagekraft von Antworten deutlich zu steigern. Kern dieses Ansatzes ist die Kombination aus einem LLM als „generativer Motor“ oder als „Herz“ und externen, nicht-parametrischen Wissensquellen, die dynamisch recherchiert und in den Antwortprozess eingebettet werden, ein Aufbau, der das Risiko von Halluzinationen senkt und die Aktualität sowie Domänenspezifität der generierten Inhalte verbessert [95]. Diese Infrastruktur versetzt LLMs in die Lage, durch intensive Vorverarbeitung umfangreicher Textkorpora menschenähnliche und zugleich fundierte Antworten zu formulieren.

Der technische Fortschritt im Bereich des NLP spiegelt sich deutlich in weit verbreiteten Sprachassistenten wie Apple Siri, Google Assistant und Amazon Alexa wider. Diese Systeme nutzen fortgeschrittene NLP-Modelle, um gesprochene Sprache effizient zu verstehen und zu verarbeiten und sind deshalb vielen Anwender*innen vertraut geworden [96]. Untersuchungen zur Nutzererfahrung bestätigen, dass Anwendungen wie Siri, Alexa und Cortana bereits fest im Alltag integriert sind und von einer breiten Nutzerschaft aktiv eingesetzt werden [97]. Damit verdeutlicht sich der unmittelbare Transfer zwischen NLP-Forschung und alltäglichem Sprachgebrauch.

Chatbots lassen sich grundsätzlich in zwei verschiedene Typen einteilen, zum einen in retrievalbasierte Systeme und generative Systeme. Retrievalbasierte Chatbots greifen auf eine Sammlung vordefinierter Antworten zurück und wählen abhängig von Eingabe und Kontext die passendste Antwort aus. Sie sind besonders zuverlässig, wenn sie mit umfangreichen und gut annotierten Datensätzen trainiert wurden, wirken jedoch oft starr und wenig menschlich. Im Gegensatz dazu erzeugen generative Chatbots mithilfe von Machine Learning und Deep Learning eigenständig neue Texte, anstatt auf feste Antwortdatenbanken zurückzugreifen. Dadurch sind sie flexibler und können komplexere, natürlichere Dialoge führen, allerdings besteht ein höheres Risiko für fehlerhafte oder unpassende Antworten. Diese Zweiteilung ist entscheidend, um die Funktionsweise, Stärken und Grenzen von

Chatbots zu verstehen und ihre Einsatzmöglichkeiten etwa im Gesundheitswesen oder in der Kundenkommunikation richtig einzuordnen. [98]

Im Gegensatz dazu erzeugen generative Chatbots ihre Antworten mithilfe tiefenlernender, selbstüberwachter Modelle. Sie sind in der Lage, neue Inhalte zu formulieren, die nicht explizit im Trainingsdatensatz vorhanden sind. Zu den derzeit leistungsfähigsten Modellen dieser Art, sogenannten State-of-the-Art (SOTA) Modellen, zählen unter anderem BERT, GPT, Gemini, Claude und LLaMA. [2]

2.5.1 Funktionsweise von RAG - Ingest

Der Ingest-Prozess bildet das Fundament jeder RAG-Architektur. Er sorgt dafür, dass externe Wissensquellen effizient und strukturiert in das System eingebracht werden, sodass LLM‘s mit aktuellen, relevanten und kontextbezogenen Informationen angereichert werden können. Die Qualität und Struktur dieses Prozesses entscheiden maßgeblich über die Leistungsfähigkeit und Zuverlässigkeit von RAG-basierten Anwendungen, insbesondere in wissensintensiven und regulierten Domänen. [99]

Zu Beginn werden Rohdaten aus unterschiedlichen Formaten (z. B. PDF, HTML, Word, Markdown) extrahiert, gesäubert und in ein einheitliches Fließtextformat überführt. Anschließend erfolgt die Segmentierung in kleinere Text-Chunks, idealerweise basierend auf einer festen Tokenanzahl oder mithilfe fortgeschrittener Algorithmen zur Erhaltung des semantischen Kontexts beitragen. [100]

Sobald die Daten entsprechend vorbereitet sind, werden die einzelnen Textabschnitte mithilfe spezialisierter Embedding-Modelle, beispielsweise BERT oder Sentence Transformers, in numerische Vektoren umgewandelt [101], [100]. Diese Embeddings erfassen die semantischen Zusammenhänge der Inhalte und ermöglichen eine effiziente Ähnlichkeitssuche innerhalb großer Datenmengen [100].

Abschließend werden die Embeddings zusammen mit den extrahierten Metadaten in einer Vektordatenbank gespeichert und indexiert [101]. Diese Datenbank bildet das Rückgrat des Retrieval-Mechanismus von RAG, sie sorgt dafür, dass relevante Informationen auch bei wachsendem Datenvolumen schnell und zuverlässig gefunden werden können.

Durch diese strukturierte und mehrstufige Verarbeitung wird sichergestellt, dass das RAG-System auf aktuelle, qualitätsgesicherte und domänenspezifische Informationen zugreifen kann, ohne dass das zugrundeliegende Sprachmodell neu trainiert werden muss. Die Forschung zeigt dabei deutlich, dass eine optimierte Ingest-Pipeline der kritische Erfolgsfaktor für die Gesamtleistung von RAG-Anwendungen ist.

2.5.2 Funktionsweise von RAG - Query

RAG kombiniert die Stärken von vortrainierten Sprachmodellen mit einem externem, nicht-parametrischem Gedächtnis, um präzise und aktuelle Antworten zu liefern. Dabei spielt der Query-Schritt eine zentrale Rolle im Zusammenspiel zwischen Benutzeranfrage, Retrieval-Komponente und Sprachmodell [67]. Zunächst wird die Benutzereingabe in eine Embedding-Repräsentation überführt, typischerweise mittels eines semantischen Encoders wie BERT-

basierten Modellen [100], [67]. Diese Vektor-Repräsentation wird im nächsten Schritt genutzt, um in einer Vektordatenbank ähnliche, relevante Dokumentabschnitte zu finden, basierend auf semantischer Nähe. Gefundenen Treffer (Retrieved Nodes) werden anschließend als kontextuelle Informationen in den Prompt des LLM integriert, um dessen Antwortqualität zu verbessern. [67]

Der Query-Schritt stellt somit sicher, dass Anfragen stets mit extern validiertem Wissen ergänzt werden können. Darüber hinaus verbessert ein gut konfigurierter Retriever sowohl Präzision als auch Geschwindigkeit der Fragenbeantwortung. [102]

Der gesamte Query-Flow ist somit ein wichtiger Bestandteil für die Gesamteffektivität und Zuverlässigkeit von RAG-Systemen. Eine klare Dokumentation und Bewertung des Query-Pfads ist in wissensintensiven Anwendungen unabdingbar. Nur so lässt sich Transparenz, Antwortqualität und Governance garantieren.

2.5.3 Einsatzgebiete von RAG im Bildungsbereich

RAG hat sich im Bildungsbereich rasch als bedeutendes Konzept etabliert, da es die Stärken großer Sprachmodelle (LLMs) mit externen Wissensquellen verknüpft und so häufig auftretende Schwachstellen wie Halluzinationen abmildert, ein Vorteil, der insbesondere in formularientlastenden Lernsettings essenziell ist [103]. RAG-basierte Chatbots ermöglichen authentische, dialogische Unterstützung, indem sie Inhalte aus Curricula, Lehrbüchern oder institutionellen Datenbanken abrufen und in Echtzeit in ihre Antworten einbauen [104].

Eine systematische Übersichtsarbeit zeigt, dass RAG-Chatbots sehr vielfältige pädagogische Funktionen übernehmen, etwa als FAQ-Assistenten, Lernhilfen, Feedbackgeber oder Tutor*innen für individualisiertes Lernen und dies durch relativ einfache Implementierung bei klar definierbarem Zweck [104]. In der Programmiersprachausbildung führte ein Einsatz von RAG zur Unterstützung beim Python-Lernen zu einer Reduktion der kognitiven Last und gesteigertem Lernverständnis, Gespräche und Erklärungen waren präziser, kontextsensitiver und halfen den Lernenden, komplexe Konzepte fundierter zu begreifen [105].

In Mathematikfragen zeigt sich ein weiterer Einsatzbereich von RAG, das Inhalte aus hochwertigen Schulbüchern nutzt und dadurch die Qualität der Antworten verbessert sowie von Lernenden tendenziell bevorzugt wird. Gleichzeitig ist jedoch ein Trade-off zwischen stark faktenbasierter und studentisch ansprechender Formulierung zu berücksichtigen[106]. Ein weiterer essenzieller Anwendungsfall ist das automatische Kurzantwort-Grading. Ein adaptiver RAG-Ansatz, der domänenspezifisches Fachwissen dynamisch ins System einbindet, führte zu deutlich höherer Bewertungsgenauigkeit gegenüber generischen LLM-Ansätzen [107].

Im Hochschulumfeld wird RAG als virtuelle Assistenz und Lehrhilfe genutzt. Eine Studie befragte Fakultätsmitglieder zu ihren Erwartungen und Bedenken bezüglich RAG-Systemen in der Informatik-Lehre. Die Lehrenden sahen großes Potenzial für virtuelle Tutor*innen und Autorenhilfen, zugleich betonten sie wichtige ethische und praxisbezogene Voraussetzungen für eine Implementierung im Unterricht. [103]

Die Untersuchung von Li et al. [60] verdeutlicht, dass RAG-Systeme vor allem in intelligenten Tutoring-Systemen, Chatbots und Empfehlungssystemen Anwendung finden. Dabei kommen zunehmend dichte und hybride Retrieval-Frameworks zum Einsatz, die im Vergleich zu

klassischen Ansätzen semantisch reichhaltigere und damit relevantere Ergebnisse liefern. In Bezug auf die Evaluationsmethoden zeigt sich, dass neben aufgaben- und domänenspezifischen Benchmarks auch Nutzerstudien zur Akzeptanz und Wirksamkeit eine zentrale Rolle spielen. Besonders betont wird die Notwendigkeit, Domänenwissen systematisch in die Retrieval- und Generierungsprozesse einzubinden, um den spezifischen Anforderungen des Bildungsbereichs gerecht zu werden. Darüber hinaus wird hervorgehoben, dass künftige RAG-Systeme stärker auf Erklärbarkeit, Fairness und pädagogische Effektivität ausgerichtet sein sollten, um langfristig eine nachhaltige Wirkung in Bildungskontexten entfalten zu können. [60]

2.5.4 Large Language Model (LLM)

Durch den rasanten Fortschritt im Bereich der Künstlichen Intelligenz haben LLMs stark an Bedeutung im Bereich der natürlichen Sprachverarbeitung gewonnen. Sie haben die Technologie in diesem Feld maßgeblich revolutioniert, indem sie es Computern ermöglichen, menschliche Sprache besser zu verstehen und zu erzeugen. Damit leisten sie einen wichtigen Beitrag zur Weiterentwicklung von KI-Anwendungen. [108]

Angesichts der hohen Rechenkosten beim Training großer Sprachmodelle besteht ein wachsendes Interesse an kleineren Modellen mit vergleichbarer Leistung. Studien zeigen, dass kleinere Modelle wie LLaMA 8B in spezialisierten Bereichen wie der Materialwissenschaft bessere Ergebnisse liefern können als größere Modelle wie LLaMA 70B. [109]

Um das Funktionsprinzip von LLMs umfassender zu verstehen, ist es wichtig, die zugrunde liegenden technischen Mechanismen zu betrachten. Zentrale Grundlage ist die Transformer-Architektur, die auf sogenannten Self-Attention-Mechanismen (SAM) basiert. Diese erlauben es dem Modell, Abhängigkeiten zwischen Wörtern und Tokens auch über lange Textpassagen hinweg zu berücksichtigen und damit kontextsensitivere Vorhersagen zu treffen. Das Training erfolgt in der Regel auf riesigen Textkorpora, wobei statistische Muster der Sprache erlernt werden. Eingaben werden dabei in Vektorrepräsentationen (Embeddings) umgewandelt, sodass semantische Beziehungen zwischen Wörtern mathematisch abgebildet werden können. Durch die wiederholte Anwendung vieler Schichten (Layer) wird der Kontext schrittweise verfeinert und so eine immer genauere Vorhersage des nächsten Tokens ermöglicht. [110]

2.5.4.1 Einsatz von Large Language Models in der Bildung

Der Einsatz von LLM‘s im Bildungsbereich eröffnet vielfältige Möglichkeiten zur Unterstützung von Lehr- und Lernprozessen. So zeigen erste systematische Literaturübersichten, dass LLM‘s häufig als virtuelle Tutor*innen eingesetzt werden, etwa für automatische Fragenproduktion, Feedbackgenerierung oder adaptive Lernhilfen, besonders populäre Modelle sind GPT und BERT [111].

Laut Sharma et al. [112] liegt dieses Potenzial darin, dass LLMs Lernumgebungen schaffen können, die individuelle Lernstile, Vorkenntnisse und Bedürfnisse berücksichtigen und so eine stärkere Personalisierung ermöglichen. Zudem können sie durch natürliche Sprachverarbeitung empathische und kontextsensitivere Antworten geben, was die emotionale Unterstützung von Lernenden stärkt. Ein weiterer Vorteil ist, dass LLMs interaktive

Rollenspiele oder simulationsbasierte Lernumgebungen gestalten können, in denen soziale und kulturelle Kompetenzen praktisch erprobt werden. Darüber hinaus erleichtern sie interkulturelle Lernsettings, indem sie unterschiedliche Perspektiven und Kontexte in Lerninteraktionen einbringen. Sharma et al. [112] betonen außerdem, dass LLMs in der Lage sind, Feedback in Echtzeit zu liefern, das nicht nur korrektiv, sondern auch motivierend wirkt. Schließlich fördern sie reflexives und dialogisches Lernen, indem sie Lernende zum Nachdenken, Fragenstellen und kritischen Diskutieren anregen. [112]

Im Bereich der Bildungsforschung lassen sich LLMs zur Analyse qualitativer Daten nutzen etwa durch automatische Auswertung von Umfrageantworten mittels fortgeschrittener Prompt-Techniken wie chain-of-thought (COT), die menschliche Leistung erreichen oder übertreffen können [113]. In der Informatik- und Ingenieurausbildung zeigen LLMs, wie in einer umfangreichen Übersicht dargelegt, Auswirkungen auf Lernende, indem sie Programmierunterstützung leisten und komplexe Konzepte verständlich aufbereiten [114]. Studien behandeln zudem die Rolle von LLMs bei der Förderung emotionaler und sozialer Kompetenzen, etwa durch Simulationen, die kulturelle Kontexte und idiomatische Ausdrücke erklären, sinnvoll besonders in Fremdsprachenunterricht oder kulturbezogener Bildung [111].

Gleichzeitig befasst sich die Literatur mit Grenzen und Risiken. Die stochastische Natur aktueller LLMs birgt die Gefahr von Halluzinationen, was kritische Reflexion und Domänenexpertise unerlässlich macht [115]. Zudem wird betont, dass eine unkritische Anwendung von LLMs produktive Lernprozesse, insbesondere productive struggle, beeinträchtigen kann, wodurch wichtige Lerngelegenheiten verloren gehen [116]. In Summe zeigen diese Studien, dass LLMs als leistungsfähige Assistenz in der Lehre wirken, sei es bei Inhaltsbereitstellung, Feedback oder emotionaler Unterstützung, dabei erfordern sie sorgfältige didaktische Integration, kritische Begleitung und ethische Reflexion.

2.5.4.2 Stärken und Schwächen von LLMs in interaktiven Kontexte

LLMs wie GPT-4, PaLM oder LLaMA haben in den letzten Jahren erhebliche Fortschritte in der Verarbeitung natürlicher Sprache erzielt und werden zunehmend in interaktiven Anwendungen eingesetzt [117].

Ein zentraler Vorteil von LLMs in interaktiven Kontexten ist ihre Fähigkeit, menschenähnliche Dialoge zu führen und auf eine Vielzahl von Anfragen flexibel zu reagieren. Studien zeigen, dass LLMs in der Lage sind, komplexe Aufgaben wie Textzusammenfassungen, Beantwortung von Fragen oder das Generieren von Programmcode effizient zu unterstützen [118]. Darüber hinaus ermöglichen Techniken wie In-Context Learning und Reinforcement Learning from Human Feedback (RLHF) eine schnelle Anpassung an spezifische Anwendungsfälle, ohne dass aufwendiges Retraining erforderlich ist [119].

In kollaborativen Szenarien fördern LLMs die Produktivität, indem sie Routineaufgaben automatisieren und als intelligente Assistenten agieren [120]. Sie bieten zudem einen niederschwelligen Zugang zu komplexen Informationssystemen, da Nutzer*innen keine Programmiersprachen oder spezielle Schnittstellenkenntnisse benötigen [121].

Trotz dieser Vorteile bestehen erhebliche Herausforderungen. Ein zentrales Problem ist die Tendenz von LLMs, sogenannte "Halluzinationen" zu produzieren, also inhaltlich falsche, aber sprachlich plausible Aussagen zu generieren [122]. Dies stellt insbesondere im

wissenschaftlichen oder sicherheitskritischen Kontext ein Risiko dar. Die mangelnde Transparenz und Nachvollziehbarkeit der Modellentscheidungen erschwert zudem die Fehleranalyse und Vertrauensbildung [121], [123].

Ein weiteres Defizit ist die fehlende Aktualität des Wissens, da LLMs auf statischen Datensätzen trainiert werden und neue Informationen nicht automatisch aufnehmen können [120]. Darüber hinaus zeigt Bender et al. [124] auf, dass LLMs bestehende gesellschaftliche Vorurteile (Bias) reproduzieren und verstärken können, was zu ethischen und sozialen Herausforderungen führt.

In interaktiven Lern- und Arbeitskontexten sind LLMs zudem oft nicht in der Lage, spezifisches, nuanciertes Feedback zu geben oder die Qualität von Nutzerbeiträgen differenziert zu bewerten [125]. Die Interaktion ist häufig prompt-abhängig, was zu Inkonsistenzen in den Antworten führen kann.

2.5.5 Embedding Modelle

Embedding-Modelle sind ein grundlegendes Element moderner KI-Systeme, da sie unstrukturierte Daten wie Wörter, Sätze oder Wissensgraphen in numerische Vektorformen überführen, in denen semantische Beziehungen messbar werden. Text-Embedding-Modelle übertragen Semantik und Syntax in hochdimensionale Räume, sodass ähnliche Konzepte nahe beieinander liegen, etwa „Groß“ und „Großartig“, womit sie Grundlage für Anwendungen wie semantische Suche oder Ähnlichkeitsmessung sind [126]. Asudani et al. [127] beschreiben Wort-Embeddings als n-dimensionale Darstellung, die Bedeutung codieren und durch Deep Learning (DL)- Modelle hergestellt werden, ein Verfahren, das Textanalysen deutlich verbessert. Kontextualisierte Modelle wie BERT heben sich ab, da sie für jedes Token kontextspezifische Repräsentationen liefern und dadurch Mehrdeutigkeiten auflösen.

Knowledge Graph Embedding (KGE) überträgt die Struktur von Wissensgraphen in Vektorräume, wobei Entitäten und Relationen abgebildet werden, damit semantische Aufgaben wie Link Prediction oder Cluster-Bildung möglich sind [127]. Moderne KGE-Modelle nutzen unterschiedliche mathematische Räume (algebraisch, geometrisch, analytisch), um vielfältige Anforderungen an Repräsentation und Performanz zu erfüllen [128].

Konzeptuell betont Aceves et al. [129], , dass Wort-Embedding-Modelle eine geometrische Interpretation ermöglichen, wodurch sich Bedeutungsverschiebungen in semantischen Räumen nachvollziehen lassen. Blagec et al. [130] konnten in ihrer Studie zeigen, dass neuronale Embeddings ohne aufwendige Ontologien eine exzellente semantische Genauigkeit im biomedizinischen Bereich erreichen (Pearson-R = 0,819). In der Praxis ist die Modellwahl entscheidend für Leistung und Interpretierbarkeit, einfache Embeddings wie Word2Vec, GloVe oder N-grams bieten effiziente Basisrepräsentationen, während tiefe Modelle (z. B. BERT-basierte) deutlich präziser, aber rechenintensiver sind [131].

Die Güte von Text-Embeddings lässt sich über Messgrößen wie Construct Validity evaluieren, zentrale Aspekte sind dabei Passung zum Task und semantische Konsistenz [126].

In Summe sind Embeddings universelle Repräsentationen, die das Brückenglied zwischen unstrukturierten Daten und modellbasierter Verarbeitung bilden. Sie ermöglichen semantische Suche, Clustering, Klassifikation, Retrieval (z. B. RAG-Kontext) und sind Kernkomponenten in

NLP, Wissensmanagement und KI-Anwendungen. Ihre Qualität beeinflusst Aussagetiefe, Effizienz und Skalierbarkeit von KI-Systemen und ihre Auswahl erfordert sorgfältige Abwägung zwischen Komplexität, Präzision und Ressourceneinsatz.

2.5.6 Retriever

Eine gründliche Aufbereitung der Datengrundlagen bildet das Fundament für leistungsstarke RAG-Systeme, da sie maßgeblich dazu beiträgt, sowohl Anfragen als auch Dokumente verständlich zu gestalten und semantisch aufeinander abzustimmen, dafür gibt es die Retrieval [132].

2.5.6.1 Bedeutung von Retrieval

Der Einsatz von Retrieval-Methoden hat sich als essenziell für die Leistungsfähigkeit moderner LLMs erwiesen. Während LLMs durch ihr parametrisiertes Wissen beeindruckende generative Fähigkeiten besitzen, sind sie in Bezug auf Aktualität, Domänenspezifik und faktische Verlässlichkeit limitiert. Retrieval-basierte Ansätze adressieren diese Schwächen, indem sie LLMs ermöglichen, während der Generierung gezielt auf externe Wissensquellen zuzugreifen und so ihre Antworten mit aktuellen und überprüfbaren Informationen anzureichern. [133], [95], [134]

Die Bedeutung von Retrieval zeigt sich besonders in wissensintensiven Aufgaben wie Fragebeantwortung, Zusammenfassung oder Entscheidungsunterstützung. Die Forschung belegt, dass RAG-Modelle die Leistung von LLMs signifikant verbessern, indem sie die Halluzinationsrate senken kann und die faktische Korrektheit erhöhen [95], [135]. Auch im Vergleich zu LLMs mit großen Kontextfenstern bleibt Retrieval ein entscheidender Faktor. Selbst bei erweiterten Kontextlängen führt die gezielte Einbindung externer Informationen durch Retrieval zu besseren Ergebnissen und effizienterer Nutzung von Ressource. [135]

Darüber hinaus ermöglicht Retrieval die Integration von proprietären, domänenspezifischen oder aktuellen Daten, ohne dass das LLM selbst neu trainiert werden muss. Dies erleichtert die Anpassung an sich schnell verändernde Wissensstände und spezifische Anwendungsbereiche. Die Forschung hebt zudem hervor, dass die Qualität des Retrievals, etwa durch adaptive oder hybride Methoden, Reranking und Kontextfilterung, maßgeblich für die Gesamtleistung des Systems ist. [134]

2.5.6.2 Diverse Retrieval Strategien

Im Zentrum jeder Retrieval-Augmented-Generation-Pipeline (RAG) steht die Auswahl geeigneter Retrieval-Strategien, um aus einer großen Wissensbasis die für eine Anfrage relevantesten Informationshäppchen zu extrahieren. Es haben sich mehrere komplementäre Verfahren etabliert, die sich anhand ihrer Suchmodalitäten und -ziele unterscheiden.

Zunächst bilden sparse Retrieval-Methoden nach dem Vorbild von BM25 den klassischen Ausgangspunkt. Sie durchsuchen den Index auf Basis von Schlüsselwortübereinstimmungen und Termgewichtungen und ermöglichen damit eine schnelle Erstselektion. Obwohl sie hohe Präzision bei eindeutig formulierten Anfragen erreichen, stoßen sie bei semantisch komplexen

oder anders formulierten Fragen an ihre Grenzen, da sie nur unzureichend Synonyme und Bedeutungsnuancen erfassen. [95]

Dense Retrieval, wie beim Dense Passage Retrieval (DPR), adressiert diese Schwäche, indem er sowohl Anfrage als auch Dokumente in dichte Vektoren überführt und sie mittels Kosinusähnlichkeit vergleicht. Durch kontrastives Training auf Query-Passage-Paaren erzielt DPR robustere semantische Treffer, benötigt jedoch aufwändigere Vektorisierungs- und Indexierungsprozesse. [95]

Um die Vorteile beider Welten zu verbinden, haben sich hybride Retrieval-Strategien durchgesetzt, die sparse und dense Scores gewichtet kombinieren. Ein typisches Vorgehen ist, zunächst jeweils eine Top-k-Liste von BM25 und DPR zu ermitteln und diese Kandidaten anschließend anhand ihrer kombinierten Scores zu mergen. Solche hybriden Ansätze steigern Recall und Precision zugleich und eignen sich besonders in Domänen mit heterogenen Dokumententypen. [95]

Für Fragestellungen, die ein breites Spektrum an thematischen Aspekten abdecken, ist die Maximum Marginal Relevance (MMR)-Methode ein bewährtes Mittel zur Sicherstellung von Diversität. MMR wählt Dokumente iterativ so aus, dass sie einerseits zur Anfrage passen, andererseits aber auch untereinander möglichst unterschiedliche Inhalte bieten. Dies verhindert Redundanzen und liefert eine ausgewogenere Ergebnisliste. [136]

Darüber hinaus erhöhen Query-Expansion-Techniken die Trefferquote, indem sie die Ursprungsanfrage um Synonyme, Oberbegriffe oder durch hypothetische Antworten (HyDE) generierte Pseudo-Queries erweitern. Dadurch lassen sich relevante Dokumente erschließen, die im Originalquery nicht explizit adressiert werden. [95]

Abschließend gewinnen hierarchische Indexierungs- und Routing-Strategien an Bedeutung. Durch den Aufbau mehrstufiger Indizes oder spezialisierter Pipelines (etwa getrennte Pfade für FAQ-Anfragen versus freitextliche Recherchen) lässt sich die Suchbreite dynamisch an die Anfragetypik anpassen, was Effizienz und Trefferqualität weiter optimiert [137].

Die Kombination und sorgfältige Abstimmung dieser Retrieval-Strategien ermöglicht es, RAG-Systeme flexibel an verschiedene Domänen, Dokumentenformate und Nutzerbedürfnisse anzupassen und so eine exzellente Balance aus Effizienz, Relevanz und Informationsvielfalt zu erzielen.

2.5.7 Reranker

Der Einsatz von Rerankern ist ein weiteres zentrales Element moderner RAG-Pipelines, da sie die Qualität der Suchergebnisse entscheidend verbessern. Während die erste Retrieval-Stufe in der Regel auf semantischen Embeddings und Approximate Nearest Neighbor (ANN)-Verfahren basiert und eine Menge potenziell relevanter Dokumente liefert, besteht die Aufgabe des Rerankers darin, diese vorselektierten Ergebnisse nach Relevanz neu zu ordnen. Reranker-Modelle nutzen hierfür häufig komplexere Sprachmodelle oder Transformer-Architekturen, die in der Lage sind, feinere semantische Unterschiede und kontextuelle Beziehungen zu erfassen als reine Vektorähnlichkeiten [138].

Besonders bewährt haben sich Cross-Encoder-Reranker (CER), bei denen Anfrage und Dokumentpassage gemeinsam in das Modell eingegeben werden. Dadurch können tiefergehende semantische Wechselwirkungen berücksichtigt werden, die über die rein numerische Nähe im Vektorraum hinausgehen [139]. In der Studie von Masanneck et al. [117] wird gezeigt, dass Reranker signifikante Verbesserungen in Aufgaben wie Question Answering, Passage Retrieval und der evidenzbasierten Antwortgenerierung erzielen. Im Gegensatz zu klassischen Dense Retrieval Methoden ermöglichen sie eine präzisere Gewichtung von Relevanzsignalen, was in wissensintensiven Domänen entscheidend ist.

Ein wichtiger Forschungsaspekt liegt zudem in der Balance zwischen Genauigkeit und Effizienz. Während CER eine sehr hohe Relevanzqualität erreichen, sind sie rechenintensiv. Daher werden in der Praxis oft zweistufige Pipelines eingesetzt, bei denen zunächst ein schneller Bi-Encoder Retriever eine Kandidatenliste generiert und anschließend ein Reranker für die Feinordnung sorgt [140]. Ergänzend kommen neural re-ranking Methoden wie MonoT5 oder ColBERTv2 zum Einsatz, die skalierbaren und dennoch akkuraten Ergebnisse liefern [117].

Darüber hinaus gewinnen domänenspezifische Reranker an Bedeutung, da sie durch Feintuning auf spezifische Fachtexte eine deutliche Leistungssteigerung erzielen können. Auch im Kontext von RAG dienen sie als wichtige Kontrollinstanz, um die Wahrscheinlichkeit von Halluzinationen zu verringern, indem sie irrelevante oder schwach kontextualisierte Passagen aus den Top-Ergebnissen entfernt werden, steigt die faktische Genauigkeit der generierten Antworten. Damit stellen Reranker ein wesentliches Qualitätsfilter in der RAG-Pipeline dar, das sowohl die Benutzerzufriedenheit als auch die Verlässlichkeit der Systeme maßgeblich beeinflusst.

2.5.8 Relevanz von Prompt Engineering

Prompt Engineering bezeichnet den gezielten Entwurf und die Optimierung von Eingabeaufforderungen für große Sprachmodelle, mit dem Ziel, präzise und relevante Ausgaben zu erzeugen [109]. Um Halluzination vorzubeugen kann man im Prompt Template „Do not use extra knowledge“ mitgeben [141]. Halluzination ist nämlich ein großes Problem von LLM’s [141], [142], [122], [109]. Sorgfältig formulierte Prompts im passenden Kontext können die Tendenz großer Sprachmodelle zu Halluzinationen deutlich verringern [109].

Prompt Engineering umfasst zudem fortschrittliche Techniken wie Chain-of-Thought oder Reflection, die das LLM dazu bringen, seine Gedankenschritte explizit auszuformulieren und somit nachvollziehbarer und weniger fehleranfällig zu agieren [143]. In der Forschung wird Prompt Engineering mittlerweile als strukturierte Disziplin betrachtet, die Taxonomien von über 50 spezifischen Techniken bietet und Best Practices für die Anwendung in modernen LLMs etwa ChatGPT definiert [144].

Die Relevanz effektiver Prompting-Strategien in lernbezogenen Anwendungskontexten von RAG-gestützten Systemen wird besonders eindrücklich durch die Studie „Understanding Learner-LLM Chatbot Interactions and the Impact of Prompting Guidelines“ von Koyuturk et al. [145] hervorgehoben. Die Autor*innen analysieren die Interaktion von Lernenden mit einem Large Language Model (LLM)-basierten Chatbot und zeigen, dass strukturierte und explizit vermittelte Prompting-Regeln einen signifikanten Einfluss auf die Qualität der generierten

Antworten haben. Dabei betrifft die Verbesserung nicht nur die inhaltliche Präzision, sondern auch die wahrgenommene Benutzerfreundlichkeit der Systeme [145].

Besonders deutlich wird der Effekt bei Nutzer*innen mit begrenzter Vorerfahrung, ohne gezielte Anleitungen formulieren, diese häufig unklare oder vage Anfragen, was zu suboptimalen Antworten führt, ein Risiko, das sich in RAG-gestützten Systemen durch das dynamische Nachladen externer Wissensquellen weiter verschärft. Die Studie belegt, dass durch gezielte Prompt-Schulung nicht nur die Nutzungsintention, sondern auch die kognitive Effizienz im Umgang mit dem Chatbot gestärkt wird. Prompt-Kompetenz fungiert dabei als eine Art „Bedienkompetenz“, die zunehmend zur Grundvoraussetzung wird, um RAG-Systeme effektiv zu nutzen [145].

2.5.9 Herausforderungen von RAG Pipelines

Trotz bereits der genannten Vorteile der RAG Pipelines stehen sie dennoch vor einer Vielzahl komplexer Herausforderungen, die sich aus der engen Verzahnung von Retrieval- und Generationskomponenten ergeben. [100]

2.5.9.1 Halluzination

Das LLM-Modell kann fehlerhafte oder ungenaue Informationen liefern, Fakten verfälschen oder Zusammenhänge suggerieren, die in der ursprünglichen Quelle nicht vorhanden sind. Solche Probleme führen oft zu irreführenden Zusammenfassungen oder logischen Widersprüchen, wenn der Kontext des Eingabe-Prompts nicht korrekt verstanden wird. Unter dem Begriff "Halluzination" versteht man die Erzeugung von Inhalten, die keinen Bezug zur Realität haben. Dabei kann das Modell überzeugend klingende, aber komplett erfundene Informationen generieren. Dieses Verhalten stellt ein zentrales Problem dar, das die Verlässlichkeit solcher Systeme infrage stellt. Hinzu kommt, dass sich Wissen ständig weiterentwickelt, während große Sprachmodelle meist auf statischen Datensätzen basieren, was zu veralteten Informationen führen kann. [141]

Wie bereits im Kapitel 2.5.8 beschrieben, ist Prompt Engineering ein zentraler Ansatz zur Reduktion von Halluzinationen in KI-Systemen. Durch gezielt formulierte Eingaben kann die Wahrscheinlichkeit fehlerhafter oder inkonsistenter Antworten verringert werden. Damit erweist sich Prompt Engineering als eine praktikable Methode, um die Zuverlässigkeit generierter Inhalte zu erhöhen.

Eine weitere Möglichkeit besteht darin, moderate Temperatureinstellungen (0,2-0,4) in RAG-Konfigurationen zu nutzen, da sie einen optimalen Kompromiss zwischen Antwortvielfalt und Faktentreue ermöglichen, während hohe Temperaturen zwar die Kreativität steigern, zugleich jedoch das Risiko von Halluzinationen deutlich erhöhen. Die Temperatur eines Large Language Models (LLM) ist ein Hyperparameter, der die Zufälligkeit der generierten Ausgaben steuert. Der Temperaturparameter eines LLM reguliert den Grad der Zufälligkeit, was zu vielfältigeren Ergebnissen führt, daher wird er oft als Kreativitätsparameter bezeichnet. [146]

2.5.9.2 Embedding- und LLM-Modellwahl

Die Wahl passender Embedding- und LLM-Modelle stellt in RAG-Pipelines eine weitere zentrale Herausforderung dar, da Modellarchitektur und Domänenanpassung entscheidend die Performance beeinflussen. Eine Studie von Oro et al. [147] zeigte, dass bei multilingualem Retrieval-Augmented Generation unterschiedliche Embeddings und Modelle wie GPT-4o, LLaMA-3.1 8B und Mistral-Nemo in Qualität und Faktenverankerung stark variieren, die Wahl beeinflusst direkt die Antwortrelevanz und Verlässlichkeit. Hinzu kommt, dass dominierende Modelle zwar hohe Relevanz liefern, aber teilweise an Fakten gebunden bleiben müssen, um Halluzinationen zu vermeiden.

Ein Artikel von Yu et al. [148] zur Modellintegration bietet einen umfassenden Leitfaden zur Einbettung unterschiedlicher LLMs in RAG-Workflows und diskutiert Auswahlkriterien wie Leistungsfähigkeit, Ressourcenbedarf und Modellkomplexität. Ergänzend begründen Yu et al. [148], dass die Modellwahl in RAG-Pipelines maßgeblich durch betriebliche Zwänge wie Datenschutz/Compliance und Kosten geprägt ist, Cloud-LLMs erhöhen das Risiko für sensible Daten und sind teuer, weshalb lokal betreibbare, ressourcenschonende Small Language Models (SLM) oft vorzuziehen sind. Zudem beschreiben sie das „Impossible Triangle“ (Genauigkeit, Kontextlänge, Antwortzeit), das harte Trade-offs erzwingt und damit die konkrete Kombination aus Embedding-Methode und SLM leitet. Praktisch empfehlen sie deshalb konsistente Embeddings (z. B. bge-base-en-v1.5 mit 768 Dimensionen) und einen performanten Vektorspeicher wie Milvus und ein instruktionsstarkes 7B-SLM (z. B. neural-chat-7b), das auf Consumer-Hardware (≈ ≥ 8 GB VRAM) läuft und so RAG-Genauigkeit bei vertretbarer Latenz ermöglicht. [148]

Sowohl Embedding-Modelle als auch LLMs können ungleiche oder voreingenommene Repräsentation transportieren, was für Bildungsanwendungen besonders kritisch ist [149]. Schließlich bleibt die Skalierbarkeit eine Herausforderung, da größere oder spezialisierte Modelle sich schwer in Ressourcen- und Kostensensitiven Bildungssystemen integrieren lassen. Die Modellwahl ist somit keine rein technische, sondern eine strategische Entscheidung, sie muss Implementierbarkeit, Leistung, Fairness und Domänenpassung in Einklang bringen.

2.5.9.3 Kontextsättigung und Chunk-Granularität

Die Abstimmung von Chunk-Granularität ist eine ebenso eine weitere Herausforderung in RAG-Pipelines, da sie das Maß an Kontext festlegt, dass einer Anfrage zugrunde liegt. Eine Studie, die zeigt, dass eine ausgewogene Chunk-Länge essenziell ist, besagt dass zu kleine Chunks den Kontext fragmentieren können, während zu große Chunks irrelevante Informationen mitliefern können. Forschungsergebnisse aus einem Multi-Dataset-Vergleich belegen, dass kurze Chunks (64-128 Tokens) besonders bei faktenspezifischen, kurzen Antworten effektiv sind, während längere Chunks (512-1024 Tokens) in dokumentübergreifenden Szenarien bessere recall-Werte liefern. Dabei variiert die optimale Chunk-Größe je nach verwendetem Embedding-Modell, zum Beispiel profitieren einige Modelle stärker vom globalen Kontext großer Chunks, andere performen besser bei hoher Detailtreue kleiner Fragmente. [150]

Weitergehende Analysen zeigen, dass contextual retrieval zwar die semantische Kohärenz besser bewahrt, jedoch mit höherem Rechenaufwand verbunden ist, während late chunking

als effizienter gilt, jedoch potenziell kontextärmer bleibt [151]. Eine zusätzliche Methode dazu ist die Mix-of-Granularity (MoG), sie zielt darauf ab, per LLM-gesteuert dynamisch die optimale Chunk-Größe je nach Anfrage zu wählen, ein Ansatz, der primär Performancegewinne durch flexible Granularitätsanpassung erzielt [152].

Diese Verfahren zeigen, dass statische Chunk-Strategien oft ineffizient sind, wenn sich Nutzerintention und Dokumentstruktur stark unterscheiden. Hinzu kommt, dass unpassende Chunk-Größen auch Latenz und Kosten beeinflussen, große Chunks erhöhen Antwortzeiten und Rechenlast. Insgesamt verdeutlicht sich, dass die Chunk-Granularität nicht nur eine technische, sondern auch eine inhaltliche Steuerungskomponente ist, die Einfluss auf Genauigkeit, Effizienz und Ressourcenverwendung im RAG-System nimmt.

2.5.9.4 Sicherheit, Datenschutz, Compliance

Sicherheit, Datenschutz und Compliance stellen weitere zentrale Herausforderungen in RAG-Pipelines dar, insbesondere in Bildungskontexten, in denen sensible Daten verarbeitet werden. Differenzielle Datenschutzverfahren wurden vorgeschlagen, um die Gefahr des Herauslesens persönlicher Informationen aus den Retrieval-Datenquellen formal zu begrenzen, ohne die Antwortqualität wesentlich zu beeinträchtigen [153]. Gleichzeitig zeigen Sicherheitsanalysen, dass RAG-Modelle potenziell weniger sicher sein können als reine LLMs, weil selbst bei sicherer Dokumentbasis unsichere Generierungen auftreten können, klassische Red-Teaming-Methoden sind in RAG-Szenarien weniger effektiv [154]. Darüber hinaus adressiert eine Studie einen lokal datenschutzfreundlichen RAG-Ansatz (LPRAG), der formale Datenschutzgarantien bietet und modellgetrieben RAG-Leistung bei eingeschränktem Datenzugriff gewährleistet [155].

Es besteht die Gefahr, dass Retrieval-Daten ungewollt sensible Informationen enthalten, Adversarial Attacks könnten durch gezielte Abfragen private Inhalte aus der Wissensbasis extrahieren, was insbesondere bei personenbezogenen Daten problematisch ist [156]. Die Erstellung einer vertrauenswürdigen RAG-Infrastruktur erfordert automatisiertes Datenschutzmanagement, Monitoring und Governance-Mechanismen, ein Ansatz, der unter dem Begriff RAGOps als erweiterte LLMOps-Strategie diskutiert, wird [157]. Angemessene Datenschutz- und Sicherheitsmaßnahmen gemäß Datenschutzprinzipien wie Privacy by Design müssen systematisch implementiert werden, um rechtliche Konformität mit Standards wie Datenschutzgrundverordnung (DSGVO) oder Bildungsregulierung zu gewährleisten.

Die Rolle von Data-Governance in LLM-Systemen, einschließlich klarer Richtlinien zur Datenqualität, Transparenz und Bias-Kontrolle ist dabei fundamental, um systemische Risiken zu minimieren und regulatorische Anforderungen zu erfüllen [158]. Schließlich muss die Modellwahl und der gesamte RAG-Betrieb durch Compliance-Checks begleitet werden, sodass sowohl technische als auch rechtliche Rahmenbedingungen konsistent eingehalten werden. Damit sind Sicherheit, Datenschutz und Compliance keine Randaspekte, sondern zentrale Querschnittsaufgaben bei der Entwicklung verlässlicher, vertrauenswürdiger RAG-Systeme im Bildungsbereich.

2.6 Interdisziplinäre Perspektiven

Die Untersuchung des Einflusses humanoider Roboter und webbasierter Chatbots mit RAG auf Akzeptanz und Benutzerfreundlichkeit im Bildungsbereich erfordert eine interdisziplinäre Herangehensweise, da die Wirkmechanismen an der Schnittstelle mehrerer Fachgebiete liegen. Fachdisziplinen wie Informatik, Psychologie, Pädagogik, Design und Soziologie tragen entscheidend dazu bei, sowohl technische Effizienz als auch soziale Akzeptanz der Systeme zu sichern [159]. Während die Chatbot-Forschung durch integrative Betrachtung von Nutzererlebnis, Plattformarchitektur und ethischen Implikationen voranschreitet, liefert die Robotik Forschung essenzielle Erkenntnisse zu Hardware-Architektur, Sensorik und Aktorik. Diese technischen Komponenten bestimmen maßgeblich das physische Design und die Interaktionsmöglichkeiten humanoider Systeme und damit ihren Gebrauchswert und ihre Glaubwürdigkeit im Bildungskontext [160]. Parallel dazu untersucht die Informatik, die Felder Maschinelles Lernen und Natural Language Processing, wie Retrieval-Augmented-Techniken Chatbots befähigen, um kontextualisiertes Wissen bereitzustellen und somit die Informationsqualität und -zuverlässigkeit zu erhöhen [161].

2.6.1 Human-Robot-Interaction (HRI)

Human-Robot Interaction (HRI) bezeichnet die interdisziplinäre Forschung und Gestaltung der Interaktion zwischen Menschen und Robotern. Dabei steht im Mittelpunkt, wie Menschen mit Robotern kommunizieren, zusammenarbeiten und diese als soziale oder funktionale Agenten erleben können. HRI integriert Erkenntnisse aus Robotik, Künstlicher Intelligenz, Kognitionswissenschaften, Human-Computer Interaction (HCI), Psychologie, Design und Verhaltensforschung. [162]

Die Zielsetzung ist es, effiziente und sichere Interaktionsmuster zu entwickeln, die den Fähigkeiten von Robotern und den menschlichen Eigenschaften gleichermaßen gerecht werden [163]. HRI-Systeme zielen meist auf Kooperation, gemeinsame Aufgabenbewältigung oder soziale Assistenz ab, bei der Menschen und Roboter andere Rollen übernehmen [162].

Die Interaktionen können verbal oder non-verbal sein, etwa über Sprache, Gestik oder Gesichtsausdruck und erfordern ein detailliertes Verständnis multimodaler Kommunikation. Insbesondere in sozialen Kontexten sind non-verbale Hinweise entscheidend, um Vertrauen, Intention und Koordination zu ermöglichen. [164]

Sicherheit ist ein zentraler Aspekt, besonders physische Sicherheit bei gemeinsamen räumlichen Aktivitäten im sogenannten Physical Human-Robot Interaction (pHRI) [165]. HRI-Systeme finden Anwendung in zahlreichen Bereichen, von industrieller Fertigung über Assistenzroboter in der Pflege bis hin zu Bildungs- und Therapieanwendungen [166].

In der Bildung fungieren soziale Roboter als Tutoren, Lernbegleiter oder Peer-Lernpartner und ermöglichen adaptives, interaktives Lernen [167]. Ein Beispiel sind humanoide Roboter, die in inklusiven Settings (Kindern mit Autismus-Spektrum-Störung (ASD) Aufmerksamkeit und Interaktion fördern [168].

Ein weiterer zentraler Aspekt ist das Vertrauen in HRI-Systeme. Nur wenn Menschen Roboter als zuverlässig, empathisch und transparent wahrnehmen, kann effektive Zusammenarbeit

gelingen [169]. Die Rolle von Vertrauen wird in der HRI-Forschung intensiv untersucht, etwa um Über- oder Untervertrauen zu vermeiden.

Methodisch adressiert HRI-Herausforderungen wie Multimodalität, soziales Verhalten, gemeinsame Absichten, adaptives Verhalten und ethische Gestaltung, häufig mithilfe von Benutzerstudien, Experimenten und Feldtests [162]. Die Forschung legt großen Wert auf Nutzendenzentrierung und sichere, menschenfreundliche Interaktion [163].

Mit zunehmenden Einsatzbereichen, von häuslicher Assistenz bis zur Bildungsrobotik, steigt auch die Bedeutung von Normen, Standardisierung und Evaluationsmethoden in HRI [163]. Insgesamt eröffnet Human-Robot Interaction vielfältige Potenziale für innovative und menschzentrierte Robotik, stellt die Forschung aber zugleich vor komplexe technische, psychologische und ethische Herausforderungen.

2.6.2 Psychologische Aspekte der Roboter-Interaktion

Die Interaktion zwischen Menschen und humanoiden Robotern stellt ein faszinierendes Forschungsfeld dar, das an der Schnittstelle zwischen Technologie und menschlicher Psychologie angesiedelt ist. Das interdisziplinäre Forschungsfeld der Robotic Psychology untersucht gezielt die emotionalen, kognitiven, sozialen und physischen Reaktionen, die menschliche Rezipienten bei Interaktion mit Robotern zeigen [170]. Dabei spielen sowohl bewusste als auch unbewusste psychologische Prozesse eine Rolle, die maßgeblich festlegen, wie Menschen Roboter wahrnehmen, mit ihnen interagieren und diese akzeptieren. Um diese Wirkmechanismen umfassend zu verstehen, bedarf es eines vertieften Dialogs zwischen experimenteller Psychologie, Sozial-Neurowissenschaften, Informatik und Robotik [171].

2.6.2.1 Anthropomorphismus und mentale Zuschreibungen

Ein zentraler psychologischer Mechanismus bei der Roboter-Interaktion ist der Anthropomorphismus, die Tendenz, nichtmenschlichen Entitäten menschliche Eigenschaften zuzuschreiben [172], [173]. Diese automatische psychologische Reaktion zeigt sich besonders deutlich in der Art, wie Menschen mentale Zustände an Roboter attribuieren. Forschungsergebnisse belegen, dass Menschen Robotern nicht nur Intentionen und Glaubensvorstellungen zuschreiben, sondern auch komplexere kognitive Fähigkeiten wie das Verstehen von Emotionen und die Fähigkeit zur Entscheidungsfindung [174]. Die Stärke dieser mentalen Zuschreibungen hängt dabei stark vom Erscheinungsbild des Roboters ab, je menschenähnlicher ein Roboter gestaltet ist, desto stärker neigen Menschen dazu, ihm mentale Eigenschaften zuzuschreiben [174].

2.6.2.2 Theory of Mind in der Mensch-Roboter-Interaktion

Die Theory of Mind (ToM), die Fähigkeit, anderen mentale Zustände wie Überzeugungen, Wünsche und Absichten zuzuschreiben, spielt eine entscheidende Rolle bei der Roboter-Interaktion [175]. Neuroimaging-Studien konnten zeigen, dass bei der Interaktion mit Robotern ähnliche Hirnregionen aktiviert werden wie bei zwischenmenschlichen Interaktionen, insbesondere der mediale präfrontale Kortex und der anteriore zinguläre Kortex [176]. Dies deutet darauf hin, dass Menschen unbewusst ihre sozialen Kognitionsmechanismen auf

Roboter anwenden. Interessanterweise führt die Wahrnehmung von ToM-Fähigkeiten bei Robotern zu einem angemesseneren Vertrauen und einer vorsichtigeren Herangehensweise in der Interaktion.

2.6.2.3 Emotionale Intelligenz und Empathie

Die emotionale Dimension der Roboter-Interaktion ist besonders bedeutsam für die Akzeptanz und Benutzerfreundlichkeit. Menschen zeigen eine bemerkenswerte Fähigkeit, empathische Reaktionen gegenüber Robotern zu entwickeln, insbesondere wenn diese emotionale Ausdrücke zeigen [177]. Studien belegen, dass die Empathie gegenüber einem Roboter dessen wahrgenommene emotionale Intelligenz signifikant erhöht [177]. Diese empathischen Reaktionen sind jedoch stark kontextabhängig und werden durch das Erscheinungsbild des Roboters, seine Bewegungen und sein Verhalten moduliert [178].

2.6.2.4 Vertrauen und Bindungsbildung

Die Entwicklung von Vertrauen in der Mensch-Roboter-Interaktion folgt spezifischen psychologischen Mechanismen, die sich von rein technischem Vertrauen unterscheiden. Beeinflusst wird dieser Prozess unter anderem durch die wahrgenommene Zuverlässigkeit des Roboters, seine anthropomorphen Eigenschaften sowie seine Fähigkeit zur sozialen Interaktion. [179]

Ullrich et al. [180] verdeutlicht in seinem Artikel, dass Menschen dazu neigen, Robotern unterschiedliche Grade von Vertrauen entgegenzubringen, abhängig von deren wahrgenommener Kompetenz und Aufrichtigkeit. Besonders interessant ist die Beobachtung, dass sowohl Unter- als auch Übervertrauen problematisch sein können, weshalb die Kalibrierung angemessener Vertrauensniveaus von entscheidender Bedeutung ist [180].

2.6.2.5 Soziale Präsenz und parasoziale Beziehungen

Ein weiterer wichtiger Aspekt ist die Entwicklung sozialer Präsenz bei Roboter-Interaktionen. Menschen können zu Robotern parasoziale Beziehungen entwickeln, einseitige emotionale Verbindungen, die typischerweise gegenüber Medienfiguren entstehen. Diese Beziehungen können durchaus positiv sein und zur Reduzierung von Einsamkeit beitragen, bergen jedoch auch Risiken der emotionalen Abhängigkeit. [181]

Die Forschung weist darauf hin, dass eine höhere wahrgenommene soziale Präsenz von Robotern die Intensität parasozialer Beziehungen verstärken kann. Langfristige Interaktionen mit Robotern zeigen, dass Nutzer*innen ihnen nicht nur funktionale Rollen zuschreiben, sondern sie zunehmend als soziale Akteure wahrnehmen. Dies eröffnet Potenziale für den Einsatz von Robotern in der Begleitung vulnerabler Gruppen, verlangt aber zugleich eine kritische Reflexion ethischer Implikationen. Besonders relevant ist die Frage, wie Designentscheidungen zur sozialen Präsenz gezielt genutzt oder reguliert werden sollten, um Abhängigkeitsrisiken zu vermeiden. [181]

Die physische Verkörperung des Roboters spielt dabei eine entscheidende Rolle, eine Studie von Jung et al. [182] belegt, dass physisch anwesende Roboter stärkere soziale Reaktionen hervorrufen als virtuelle Agenten.

2.6.2.6 Persönlichkeit und individuelle Unterschiede

Die Persönlichkeit der Nutzer*innen beeinflusst maßgeblich die Roboter-Interaktion. Menschen mit höheren Werten in Offenheit, Extraversion und Verträglichkeit zeigen eine größere Akzeptanz gegenüber Robotern. Extrovertierte Personen neigen dazu, Roboter stärker zu anthropomorphisieren, während emotional instabilere Individuen menschenähnliche Roboter als bedrohlicher empfinden. [183]

Persönlichkeitsmerkmale wirken somit als Moderatoren in der MRI und beeinflussen, ob Nutzer die Technologie als hilfreich oder als belastend erleben. Während extrovertierte und offene Personen häufiger positive Emotionen in der Interaktion berichten, reagieren introvertierte oder neurotische Nutzer eher skeptisch und zurückhaltend. Diese Unterschiede verdeutlichen, dass die Akzeptanz sozialer Roboter nicht allein durch deren technisches Design, sondern auch durch individuelle Dispositionen geprägt wird. Für die Gestaltung von Bildungsrobotern bedeutet dies, dass adaptive Interaktionsstrategien notwendig sind, die unterschiedliche Persönlichkeitstypen berücksichtigen. Ein stärker personalisiertes Design könnte dadurch die Nutzungsbereitschaft und die langfristige Integration von Robotern in Lernumgebungen erhöhen. [183]

2.7 Bewertung der Technologieakzeptanz

Mit dem zunehmenden Einfluss digitaler Technologien in nahezu allen gesellschaftlichen und wirtschaftlichen Bereichen stellt sich die Frage, warum bestimmte Technologien von Nutzer*innen angenommen werden, während andere trotz technischer Reife und potenzieller Vorteile auf Ablehnung stoßen. Die Beantwortung dieser Frage ist insbesondere für die erfolgreiche Implementierung technologischer Systeme in Organisationen von großer Bedeutung. Ein fundiertes Verständnis der zugrunde liegenden Akzeptanzmechanismen ermöglicht es, gezielte Maßnahmen zur Förderung der Nutzung zu entwickeln und somit den Nutzen digitaler Innovationen zu maximieren.

Die Bewertung der Technologieakzeptanz dient dazu, um den Erfolg neuer Systeme und Anwendungen einschätzen zu können. Sie untersucht, inwieweit Anwender*innen bereit sind, eine Technologie in ihren Alltag oder Arbeitsprozess zu integrieren. Dabei werden sowohl individuelle Einstellungen als auch äußere Einflussfaktoren berücksichtigt. Durch diese Analyse lassen sich Potenziale und Hürden frühzeitig erkennen und gezielt adressieren.

2.7.1 Begriff und Bedeutung von Technologieakzeptanz

Technologieakzeptanz beschreibt den Grad, in den Personen bereit sind, eine bestimmte Technologie freiwillig zu nutzen, basierend auf zwei zentralen Wahrnehmungen. Perceived Usefulness (PU), die Überzeugung, dass die Technologienutzung die Zielerreichung verbessert, sowie Perceived Ease of Use (PEOU), die Einschätzung, dass die Nutzung mit geringem Aufwand verbunden ist. Das Technology Acceptance Model (TAM) erklärt, dass diese Konstrukte Einstellungen zur Technologie und damit die Nutzungsabsicht beeinflussen, die wiederum das tatsächliche Verhalten steuert. [184]

2.7.2 Relevante Akzeptanzmodelle

Um die Verhaltensfaktoren zu erklären, die über die Akzeptanz moderner Technologien entscheiden, sind verschiedene theoretische Modelle entstanden und angewendet worden. Dazu zählen etwa die Theorie der Theory of Planned Behavior (TPB) [185], das Technology-Acceptance-Model [186] sowie das Unified Theory of Acceptance and Use of Technology-Modell [186].

2.7.2.1 Technology Acceptance Model

Das Technology Acceptance Model wurde von Fred D. Davis et al. [186] im Kontext betrieblicher Informationssysteme entwickelt. Es ist in der Theory of Reasoned Action (TRA) [187] und ihrer Weiterentwicklung, der TPB [185], verankert. Während TRA/TPB generelle Verhaltensabsichten (VA) erklären, überträgt TAM diesen Ansatz auf die Nutzung von Informationstechnologien und identifiziert zwei zentrale kognitiv-affektive Determinanten.

1. Perceived Usefulness: die subjektive Einschätzung, dass der Einsatz einer Technologie die eigene Arbeits- bzw. Lernleistung verbessert.

2. Perceived Ease of Use: die subjektive Wahrnehmung, dass der Umgang mit der Technologie mühelos bzw. ohne großen Aufwand möglich ist.

Davis et al. [186] postuliert, dass PEOU PU positiv beeinflusst (je leichter etwas zu bedienen ist, desto nützlicher erscheint es) und beide Konstrukte die Nutzungsabsicht (Behavioral Intention, BI) determinieren, die wiederum das tatsächliche Nutzungsverhalten (Actual Use) vorhersagt.

Während TAM den Vorteil der Einfachheit bietet, liegt ein Kritikpunkt darin, dass es den Fokus stark auf individuelle Wahrnehmungen legt. Externe Variablen wie organisationale Rahmenbedingungen, kulturelle Unterschiede oder soziale Dynamiken werden nur am Rande oder in Erweiterungen des Modells berücksichtigt. Dies führte zu verschiedenen Erweiterungen: [188]

- TAM 2 [18] ergänzt soziale Einflüsse (image, subjektive Norm) und kognitive Instrumentalprozesse (Job Relevance, Output Quality).

- TAM 3 [18] integriert Computer Self-Efficacy, Anxiety und Perceptions of External Control und spezifiziert Determinanten von PEOU detaillierter.

- Das UTAUT [18] verschmilzt TAM mit acht konkurrierenden Modellen und führt Performance Expectancy, Effort Expectancy, Social Influence und Facilitating Conditions als Kerndeterminanten ein.

Trotz theoretischer Kritik hat sich TAM in der Praxis als äußerst nützliches Werkzeug etabliert, insbesondere in der frühen Evaluierung neuer Technologien. Unternehmen und Organisationen nutzen das Modell, um abzuschätzen, ob ein System potenziell Akzeptanz finden wird, und um gezielt Maßnahmen zur Steigerung der Benutzerfreundlichkeit oder des wahrgenommenen Nutzens zu entwickeln. [188]

2.7.2.2 Unified Theory of Acceptance and Use of Technology

Das UTAUT wurde von Venkatesh, Morris, Davis und Davis [18] entwickelt, um die zersplitterte Forschung zu Technologieakzeptanz zusammenzuführen. Die Autor*innen integrierten acht etablierte Modelle (u. a. TAM, TPB, Innovation Diffusion Theory) zu einem gemeinsamen Rahmen.

Die Vielzahl der in der Vergangenheit entwickelten Modelle brachte jeweils unterschiedliche Einflussfaktoren hervor, die in einzelnen Kontexten eine hohe Erklärungskraft aufwiesen, jedoch in ihrer isolierten Anwendung oft keine allgemeingültigen Aussagen zur Technologieakzeptanz zuließen.

Das Ziel von UTAUT ist es, ein theoretisch kohärentes und zugleich empirisch robustes Modell zu schaffen, das die Varianz in der Verhaltensabsicht zur Nutzung (behavioral intention) sowie im tatsächlichen Nutzungsverhalten (use behavior) möglichst umfassend erklären kann. Die Autoren identifizierten in ihrer Metaanalyse vier zentrale Konstrukte und einem abhängigen Konstrukt, die als Hauptdeterminanten der IT-Akzeptanz fungieren (vgl. Tabelle 1).

Konstrukt	Definition
Leistungserwartung/ LE	Bezeichnet den Grad, zu dem eine Person glaubt, dass die Nutzung einer Technologie ihre Arbeitsleistung verbessern wird. Dieser Konstruktelement ist konzeptuell eng verwandt mit den Konzepten „perceived usefulness“ im TAM und „relative advantage“ in der IDT. Empirisch zeigt sich Performance Expectancy als der stärkste Prädiktor für die Nutzungsabsicht.
Aufwandserwartung / AE	Beschreibt die Leichtigkeit, mit der die jeweilige Technologie bedient werden kann. Es basiert unter anderem auf dem Konstrukt der „perceived ease of use“ im TAM sowie auf Aspekten der Komplexität im IDT. Besonders in frühen Nutzungsphasen hat Effort Expectancy einen signifikanten Einfluss auf die Intention zur Nutzung.
Sozialer Einfluss / SE	Reflektiert den wahrgenommenen sozialen Druck bzw. die Überzeugung, dass relevante Personen, wie KollegInnen, Vorgesetzte oder das soziale Umfeld erwarten, dass eine bestimmte Technologie genutzt wird. Dieser Faktor ist insbesondere in organisationalen Kontexten und bei freiwilliger Nutzung von Bedeutung.
Erleichternde Bedingungen / EB	Bezieht sich auf die wahrgenommenen organisatorischen und technischen Voraussetzungen, die die Nutzung einer Technologie erleichtern. Dazu zählen unter anderem Verfügbarkeit von Support, Ressourcen, Infrastrukturen oder Trainingsangeboten. Dieses Konstrukt beeinflusst direkt das tatsächliche Nutzungsverhalten, insbesondere bei erfahreneren Nutzer*innen.
Verhaltensabsicht / VA	Bezieht sich auf die individuelle Absicht einer Person, eine bestimmte Technologie zukünftig zu nutzen. Sie spiegelt die persönliche Einstellung zur Anwendung wider und gilt als zentraler Prädiktor für das tatsächliche Nutzungsverhalten. Eine hohe Verhaltensabsicht zeigt sich, wenn Nutzer*innen davon überzeugt sind, dass die Technologie nützlich, einfach zu bedienen und im eigenen Kontext sinnvoll einsetzbar ist. Im UTAUT-Modell wird die Verhaltensabsicht maßgeblich durch Leistungserwartung, Aufwandserwartung, sozialer Einfluss und erleichternde Bedingungen bestimmt.

Tabelle 1 - UTAUT Konstrukte nach David et. al [18]

Die Wirkung der einzelnen Pfade im UTAUT-Modell ist nicht für alle Anwender*innen in gleichem Maße ausgeprägt, da vier Moderatoren berücksichtigt werden. So zeigt sich, dass das Geschlecht einen Einfluss hat, Leistungserwartung wirkt stärker bei Männern, während Aufwandserwartung bei Frauen eine größere Rolle spielt. Auch das Alter beeinflusst die Zusammenhänge, mit zunehmendem Alter nimmt der Effekt der Leistungserwartung ab, während der Einfluss der anreizbasierten Erwartung und der Erfahrungserwartung steigt. Die Erfahrung der Nutzer*innen stellt einen weiteren Moderator dar, da mit wachsender Nutzungserfahrung die Bedeutung der Anstrengungserwartung abnimmt, während die Rolle der Erwartung an unterstützendes Verhalten zunimmt. Schließlich wirkt sich auch die Freiwilligkeit der Nutzung aus, Subjektive Normen beziehungsweise sozialer Einfluss sind bedeutsamer, wenn die Einführung einer Technologie verpflichtend erfolgt. Diese Moderationen ermöglichen differenzierte Prognosen, ohne das Modell unnötig zu verkomplizieren. [18]

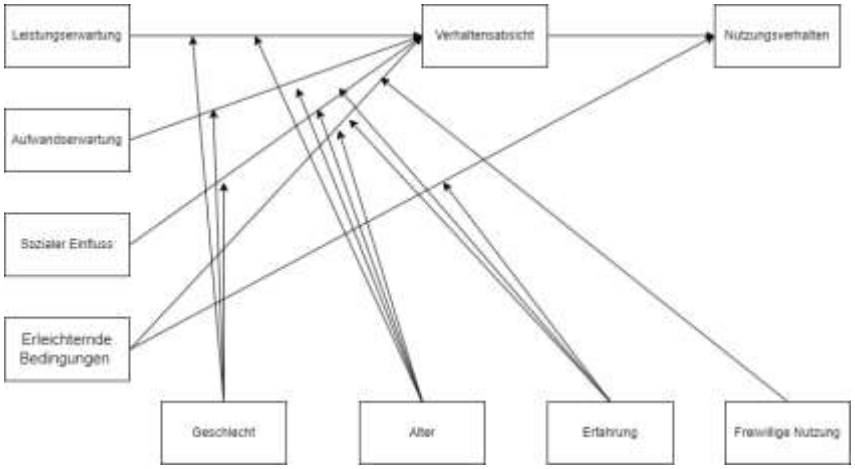

Abbildung 1 - UTAUT nach Venkatesh et al. [18]

Die empirische Validierung von UTAUT durch Venkatesh et al. [18] erfolgte anhand von vier Längsschnittstudien in realen Unternehmenskontexten, bei denen neue IT-Systeme eingeführt wurden. Die Studien umfassten insgesamt sieben Organisationen und über 1.600 Proband*innen. Die Ergebnisse zeigen, dass das UTAUT-Modell in der Lage ist, bis zu 70 % der Varianz in der Verhaltensabsicht zur Nutzung von Technologie zu erklären. Dies stellt eine signifikante Verbesserung gegenüber früheren Modellen dar, deren Erklärungskraft typischerweise zwischen 17 % und 53 % lag. [18]

2.7.2.3 UTAUT2

Während UTAUT vor allem das Nutzungsverhalten von Beschäftigten in organisationsgebundenen Pflichtsituationen erklärt, rückt UTAUT 2 den alltäglichen Umgang von Konsument*innen mit frei wählbaren Technologien ins Zentrum. Venkatesh et al. [189] ergänzten dazu drei Konstrukte, die im Consumer-Kontext besonders prägend sind (vgl. Tabelle 2).

Neues Konstrukt	Definition
Hedonistische Motivation (HM)	intrinsischer Spaß, der aus der Nutzung selbst entsteht
Preis Wert (PW)	wahrgenommene Wirtschaftlichkeit, Nutzen relativ zu monetären und nicht-monetären Kosten
Gewohnheit (GH)	automatisiertes Verhalten aufgrund wiederholter Nutzung

Tabelle 2 - UTAUT2 nach Venkatesh et al. [189]

Damit wächst das Modell von vier auf sieben direkte Prädiktoren der Verhaltensabsicht (LE, AE, SE, GH, HM, PW, HT) sowie zwei direkte Determinanten des tatsächlichen Verhaltens (GW, EB, VA). Die bekannten Moderatorvariablen, Geschlecht, Alter und Erfahrung bleiben bestehen. [189]

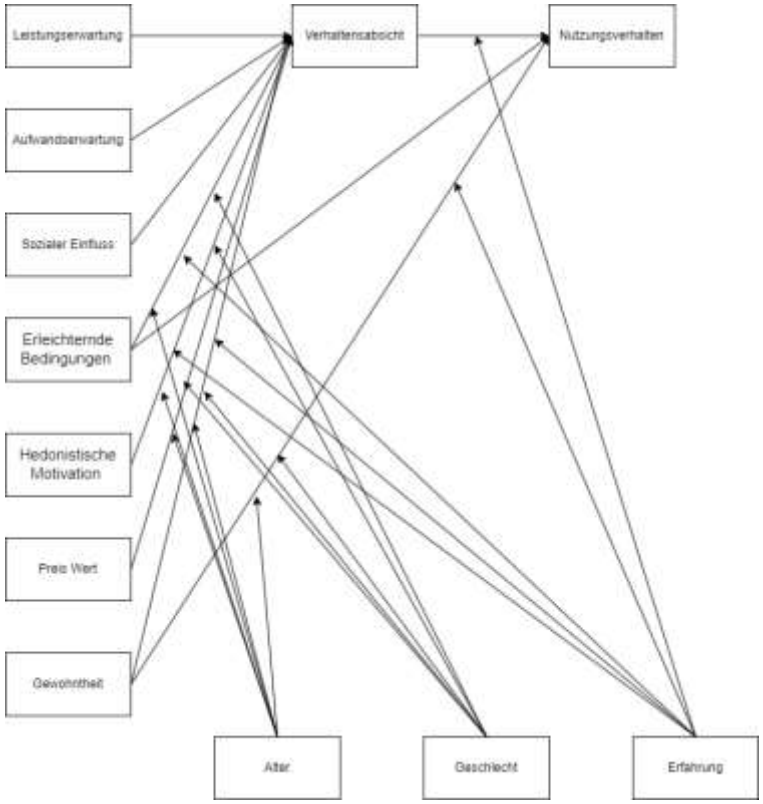

Abbildung 2 - UTAUT 2 nach Venkatesh et al. [189]

2.8 Bewertung der Benutzerfreundlichkeit

Die Bewertung der Benutzerfreundlichkeit dient dazu, die Qualität der Interaktion zwischen Menschen und Systeme zu erfassen. Im Mittelpunkt steht dabei, wie einfach, effizient und zufriedenstellend eine Anwendung genutzt werden kann. Eine hohe Benutzerfreundlichkeit trägt wesentlich zur Akzeptanz und nachhaltigen Nutzung einer Technologie bei. Durch ihre Analyse lassen sich Schwachstellen identifizieren und gezielte Verbesserungen ableiten.

2.8.1 Begriffsklärung der Benutzerfreundlichkeit / Usability

Benutzerfreundlichkeit, im Englischen Usability, bezeichnet nach ISO 9241-11 [190] das Ausmaß, in dem ein Produkt von bestimmten Nutzer*innen unter definierten Nutzungsbedingungen effektiv, effizient und zufriedenstellend genutzt werden kann, um festgelegte Ziele zu erreichen. Dabei werden drei zentrale Dimensionen unterschieden. Unter Effektivität versteht man die Genauigkeit und Vollständigkeit, mit der Nutzer*innen ihre Ziele erreichen. Effizienz beschreibt den Aufwand, beispielsweise in Form von Zeit oder kognitiven Ressourcen, den Anwender*innen zum Erreichen ihrer Ziele benötigen. Die Zufriedenheit wiederum umfasst die subjektiven Empfindungen der Anwender*innen hinsichtlich Komfort und Akzeptanz bei der Nutzung. Diese drei Kriterien bilden die Grundlage für die systematische Evaluation interaktiver Systeme und ermöglichen Vergleiche zwischen unterschiedlichen Anwendungen oder Iterationsständen im Entwicklungsprozess.

Ergänzend dazu fasst Nielsen [191] die Benutzerfreundlichkeit in fünf Qualitätsfaktoren zusammen. Dazu gehören die Learnability, also wie leicht neue Anwender*innen grundlegende Aufgaben erlernen können, sowie die Efficiency, die angibt, wie schnell erfahrene Nutzer*innen Aufgaben erledigen. Die Memorability beschreibt, wie leicht Gelegenheitsnutzer*innen nach einer Zeit der Nichtnutzung wieder in das System einsteigen können. Hinzu kommt der Faktor Errors, der sowohl die Anzahl und Schwere als auch die Erholungsfähigkeit von Bedienfehlern umfasst. Schließlich spielt die Satisfaction, also die wahrgenommene Angenehmheit der Nutzung, eine wesentliche Rolle. [191]

2.8.2 System Usability Scale

Die System Usability Scale (SUS) ist ein standardisierter Fragebogen zur Messung der Gebrauchstauglichkeit eines Systems und wurde 1986 von John Brooke entwickelt [19]. Sie besteht aus zehn Items mit einer fünfstufiger Likert-Skala, die jeweils abwechselnd positiv und negativ formulierte Aussagen enthalten. Durch Anwendung einer festen Berechnungsformel ergibt sich aus den Antworten ein globaler Score zwischen 0 und 100, wobei Werte über 68 als überdurchschnittlich gelten. SUS ist technologieunabhängig, einfach anzuwenden und eignet sich sowohl für Software als auch für Hardware, Webanwendungen und mobile Systeme. [192]

Eine Retrospektive-Analyse im Journal of Usability Studies unterstreicht die Zuverlässigkeit der SUS als schnelles, kosteneffizientes Werkzeug zur Usability-Bewertung auch außerhalb formeller Testsettings [193].

Durch seine Einfachheit und breite Anwendbarkeit ist die SUS besonders populär in Usability-Evaluierungen im Bildungs- und E-Learning-Feld. Eine Studie, die das Technology Acceptance Model (TAM) um die Einführung des SUS erweiterte, zeigt, dass wahrgenommene Benutzerfreundlichkeit signifikant die Nutzungseignung von E-Learning-Systemen beeinflusst. Dabei wurde deutlich, dass Faktoren wie soziale Normen, Systemzugang und Selbstwirksamkeit entscheidend zur Verhaltensabsicht beitragen. [194]

Die System Usability Scale bietet den Vorteil, dass sie Benutzerfreundlichkeit sowohl effektiv (Zielerreichung), effizient (Ressourceneinsatz) als auch zufriedenstellend (User Experience)

abbildet [193]. Jedoch bleibt die SUS eine generelle Kennzahl und sagt nicht explizit, welche Usability-Probleme im Detail vorliegen [195].

Zusammenfassend stellt die System Usability Scale ein bewährtes und praktikables Instrument zur quantitativen Erhebung der Benutzerfreundlichkeit dar. In Kombination mit anderen Methoden, etwa zur Akzeptanzmessung (wie UTAUT), kann sie helfen, ein umfassendes Bild der Nutzerperspektive zu gewinnen und gezielte Verbesserungspotenziale in der Gestaltung digitaler Systeme zu identifizieren.

2.9 Stärken und Schwächen von webbasierten Chatbots im Vergleich zu humanoiden Robotern

Webbasierte Chatbots lassen sich sehr schnell und skalierbar in Lernumgebungen integrieren und decken ein breites Spektrum an Fächern ab systematische Übersichten berichten, dass die meisten Bildungs-Chatbots auf Webplattformen laufen und primär lehrorientierte Funktionen übernehmen [57], [196]. Sie sind 24/7 verfügbar, kostengünstig auszurollen und entlasten Lehrende bei Routinefragen, was in Reviews als zentraler Mehrwert hervorgehoben wird [197]. Eine Meta-Analyse von Laun et.al. deutet zudem auf kleine bis moderate positive Effekte auf Lernleistungen hin, bei gleichzeitig guter Eignung für Hausaufgabenhilfe und personalisierte Unterstützung [198]. Ein wichtiger didaktischer Vorteil sind, dass Chatbots Aspekte des selbstregulierten Lernens (SRL) unterstützen können, insbesondere Ressourcenfinden, Strategien anstoßen und das eigene Lernen überwachen [199]. Gleichzeitig zeigen die SRL-Befunde Grenzen, wie Zielsetzungs-, Reflexions- und langfristige Planungsphasen werden bislang seltener adressiert, was die Wirksamkeit über längere Lernzyklen begrenzen kann [199]. Evidenzqualität und Evaluationsdesigns können variieren daher ist die Übertragbarkeit von Ergebnissen auf unterschiedliche Kontexte nicht immer gesichert [57].

Humanoide Roboter bringen dagegen verkörperte Interaktionsmöglichkeiten (Gestik, Blickverhalten, räumliche Präsenz) ein, die Motivation, Aufmerksamkeit und Involvement von Lernenden fördern können [47]. In schulischen Langzeitstudien berichten Lehrkräfte von Chancen für Engagement und Curriculumsanbindung, sofern der Einsatz didaktisch geplant ist [47]. Humanoide Roboter können als didaktische oder soziale Partner auftreten und kollaborative Lernsettings (z. B. gemeinsames Spielen/Üben) unterstützen [90]. Eine Studie von NAO als Interface zeigen, dass die physische Präsenz als Motivator wirkt und komplexe Inhalte (z. B. numerische Methoden) kontextualisiert werden können [81]. Gleichzeitig erfordern humanoide Systeme deutlich mehr Infrastruktur, Wartung und Betreuung Lehrkräfte nennen technische Robustheit und Vorbereitungsaufwand als Hürden [47]. Eine zusätzliche Einschränkung ist das Risiko der Ablenkung bzw. erhöhter kognitiver Last, wenn Aufmerksamkeit zu stark auf die Technik anstatt auf den Lerngegenstand gelenkt wird [90].

Im direkten Vergleich punkten Web-Chatbots bei Skalierbarkeit, Kosten-Effizienz, Niedrigschwelligkeit und bei asynchronen Unterstützungsszenarien [196], [57]. Sie eignen sich besonders für FAQ-artige Hilfen, formative Unterstützung und personalisierte Hinweise in großen Kohorten, mit nachweisbaren, wenn auch meist kleinen bis moderaten Effekten [198]. Ihre Schwächen liegen in begrenzter Langzeit-SRL-Abdeckung, potenziellen

Qualitäts-/Verlässlichkeitsfragen und geringerer sozialer Präsenz durch fehlende Verkörperung [199], [57]. Humanoide Roboter überzeugen, wenn sozial-interaktive, haptische oder räumliche Lernanteile im Vordergrund stehen, etwa bei jüngeren Lernenden oder in Settings, die von Gestik und geteiltem Aufmerksamkeitsfokus profitieren [90], [47]. Ihre Schwächen sind Kosten, technische Komplexität, Integrationsaufwand und mitunter uneinheitliche Evidenz zur nachhaltigen Lernwirkung quer über Fächer und Jahrgänge [47]. Nach dem Design muss bei humanoiden Systemen außerdem auf wahrnehmungspsychologische Effekte wie das Uncanny-Valley geachtet werden.

Für die Praxis ergibt sich eine komplementäre Perspektive, da Web-Chatbots skaliert zur Basissupport- und SRL-Begleitung beitragen, während humanoide Roboter in gut geplanten Präsenzszenarien soziale Präsenz und Interaktionstiefe einbringen [57]. Kombinierte Modelle z. B. Chatbot-gestützte Vor-/Nachbereitung plus verkörperte Aktivitäten im Unterricht, adressieren unterschiedliche Lernmechanismen und können Stärken beider Ansätze bündeln [57]. Entscheidend bleibt die didaktische Orchestrierung (Ziel-Aufgaben-Passung, klare Rollenverteilung, Evaluationsdesign), damit Technologieeinsatz nicht Selbstzweck wird [47]. In Summe sollten Institutionen die Wahl zwischen Web-Chatbot und humanoidem Roboter nicht als Entweder-oder verstehen, sondern entlang der Lernziele, Ressourcen und Zielgruppe treffen, evidenzbasiert und iterativ evaluiert [57].

2.9.1 Benutzerfreundlichkeit und Akzeptanz

Im Bildungsbereich zeichnen sich webbasierte Chatbots durch ihre hohe Zugänglichkeit und intuitive Bedienbarkeit aus, Studierende können über gängige Endgeräte jederzeit und ortsunabhängig interagieren, was die Nutzerfreundlichkeit erheblich erhöht und zu einer positiven Wahrnehmung beiträgt. Die dialogische Struktur von Chatbots fördert zudem die Aktivierung kognitiver Prozesse, da Lernende direkte Rückmeldungen erhalten und bei Unklarheiten zielgerichtet nachfragen können, was die Akzeptanz weiter erhöht. [200]

Die Lebendigkeit der Robotergesten und -bewegungen kann zwar die Aufmerksamkeit fördern, führt aber bei einigen Nutzer*innen zu Unsicherheit, wenn die Kommunikation zu stark anthropomorph wirkt und unerwartete Verhaltensweisen zeigt. Während Chatbots in Studien durchgehend als niederschwelliges und zuverlässiges Werkzeug bewertet werden, hängt die Akzeptanz von Robotern stark von deren Gestaltung und Verlässlichkeit ab klare, reproduzierbare Verhaltensmuster und freundlich wahrgenommene Gesten erhöhen hier deutlich die Nutzerzufriedenheit. [201]

Der Einsatz von Chatbots bietet Institutionen erhebliche Vorteile hinsichtlich Skalierbarkeit, Kosteneffizienz und Verfügbarkeit. Während menschliche Mitarbeiter zeitlich begrenzt und personalintensiv arbeiten, können Chatbots praktisch unbegrenzt skaliert werden und ermöglichen dadurch eine durchgehende 24/7-Betreuung. Studien zeigen, dass Chatbots insbesondere im Kundenservice und im Gesundheitswesen aufgrund ihrer niedrigen Betriebskosten, ihrer leichten Integration in bestehende Plattformen sowie ihrer dauerhaften Verfügbarkeit erhebliche Effizienzgewinne schaffen. [202], [203], [204], [205]

2.9.2 Emotionale und soziale Aspekte

Webbasierte Chatbots und humanoide Roboter dienen zwar beide der unterstützenden Interaktion für Anwender*innen, sie unterscheiden sich jedoch grundlegend in ihrer emotionalen Ausdrucksfähigkeit, sozialen Präsenz und Wahrnehmung durch die Anwender*innen. Web-Chatbots müssen emotionale Signale allein über text- oder sprachbasierte Kanäle vermitteln. Sie nutzen NLP und Sentiment-Erkennung, um etwa durch Wortwahl, Emoticons oder Sprachmodulation empathisch zu reagieren. Untersuchungen zeigen, dass Chatbots, die ihre „eigenen" Emotionen offenlegen, etwa Trost oder Mitgefühl ausdrücken, die Nutzerzufriedenheit und Wiederverwendungsabsicht signifikant steigern können [206]. Diese künstliche „Offenbarung" verbessert die wahrgenommene Intimität und fördert eine emotionale Bindung, bleibt jedoch auf den Dialogkanal beschränkt und kann nonverbale Aspekte nicht vermitteln.

Im Gegensatz dazu verfügen humanoide Roboter über einen physischen Körper, der nonverbale Ausdrucksmittel wie Mimik, Gestik, Blickkontakt und Proxemik einsetzt. Bereits frühere Studien im HRI-Bereich belegen, dass adaptive emotionale Ausdrucksmodelle bei Robotern wie NAO oder Cozmo dazu führen, dass Kinder und Erwachsene emotional stärker reagieren und positiver interagieren als mit rein textbasierten Agenten [207], [208]. Durch synchronisierte Körperhaltungen, Stimme und Gestik kann ein Roboter Gefühle glaubwürdiger vermitteln und so „soziale Präsenz" erzeugen, die weit über rein digitalen Interaktionen hinausreicht.

Die soziale Wahrnehmung von Web-Chatbots unterscheidet sich deutlich von physischen KI-Agenten, Chatbots wirken in ihrer Kommunikation funktional und nutzenorientiert, und ihre Empathie wird kognitiv anerkannt, jedoch meist abstrakt erlebt, da sie nur geringe soziale Hinweisreize bieten [209]. Im Gegensatz dazu zeigen humanoide Roboter durch ihre physische Präsenz eine stärkere emotionale Resonanz, teilweise vermittelt über aktivierte Spiegelneuronensysteme, was zu intensivierter sozial-emotionaler Bindung führt. [210]

Beide Ansätze profitieren von multimodaler Erweiterung, während Chatbots durch Integration von Emojis, Audio- und Videokomponenten ihre Emotionsvermittlung verbessern können, unterstützen Roboter durch Grafiken oder Hologramme eine noch reichhaltigere Interaktion. Zusammenfassend lässt sich sagen, dass webbasierte Chatbots durch gezielte emotionale Offenlegung und sentimentgesteuerte Antworten eine hohe Effizienz und Nutzerakzeptanz erreichen, während humanoide Roboter durch körperliche Präsenz und nonverbale Ausdrucksmittel eine intensivere soziale und emotionale Bindung ermöglichen. Die Wahl zwischen beiden Systemen hängt somit von Anwendungsfall, technischer Machbarkeit und gewünschtem Grad an sozialer Präsenz ab.

3 Experiment

Um die Forschungsfrage zu beantworten, wie die Verwendung eines humanoiden Roboters im Vergleich zu einem webbasierten Chatbot mit RAG die Akzeptanz und die Benutzerfreundlichkeit im Bildungsbereich beeinflusst, wird in diesem Kapitel das durchgeführte Experiment im Detail beschrieben. Zunächst wird die methodische Vorgehensweise erläutert, gefolgt von der Darstellung der Versuchsgruppen und der eingesetzten Dokumente für die RAG-Pipeline. Darauf aufbauend wird die Architektur des Systems sowie die verwendeten Technologien vorgestellt. Abschließend werden die angewandten Verfahren der statistischen Analyse beschrieben, mit denen die Ergebnisse ausgewertet und interpretiert werden.

3.1 Methodik

Es wurde vor der Stichprobenerhebung ein Testlauf durchgeführt. Dabei nahm jeweils eine Person für beide Systeme, den humanoiden Roboter NAO und den Web-Chatbot teil. Diese Teilnehmenden wurden nicht in die Hauptstudie übernommen, um die Vergleichbarkeit der späteren Ergebnisse sicherzustellen. Der Vortest diente in erster Linie dazu, mögliche Schwierigkeiten im Ablauf des Skripts zu identifizieren und zu korrigieren. Durch das Feedback der Proband*innen konnten Schwachstellen erkannt und entsprechende Anpassungen vorgenommen werden, sodass ein reibungsloser Verlauf der Haupterhebung gewährleistet war und zuverlässige Resultate erzielt werden konnten. Zusätzlich wurde auch der eingesetzte Fragebogen erprobt, um sicherzustellen, dass die Formulierungen verständlich sind und von den Befragten korrekt interpretiert werden.

3.1.1 Forschungsdesign

Zur Untersuchung der Wahrnehmung und Akzeptanz der Nutzung wurde ein experimentelles Forschungsdesign mit pre- und posttestbasierter Struktur entwickelt. Ziel war es, Unterschiede in der Interaktion und Bewertung eines webbasierten Chatbots im Vergleich zu einem humanoiden Roboter (NAO) systematisch zu analysieren. Zu Beginn erhielten alle Teilnehmenden eine standardisierte Einführung in das Untersuchungssetting, die mittels einer kurzen Präsentation vor dem Studiengang erfolgte. Darin wurden Ablauf, Zielsetzung und Datenschutzbestimmungen transparent dargestellt. Vor der Durchführung des eigentlichen Experiments mussten die Teilnehmenden eine schriftliche Einwilligung zur Erhebung und Verarbeitung ihrer Daten abgeben. Erst nach dieser Einwilligung wurde die Teilnahme am Experiment ermöglicht. Im Anschluss daran wurden allgemeine demografische Angaben sowie Vertrautheit und Nutzung mit künstlicher Intelligenz mittels eines kurzen Fragebogens erhoben, um etwaige Einflussfaktoren bei der späteren Analyse evaluieren zu können.

Der Untersuchungsablauf gliederte sich in folgende Phasen.

1. Pre-Test (PT): Erhebung des Ausgangszustands der Teilnehmenden in Bezug auf Vorkenntnisse vor der Interaktion.

2. Interaktion mit dem System: Anschließend erfolgte die Interaktion mit einem der beiden Systeme, entweder dem Web-Chatbot oder dem humanoiden Roboter NAO. Die Zuweisung erfolgte randomisiert.

3. UTAUT-Fragebogen: Direkt nach der Interaktion wurde der UTAUT-Fragebogen ausgefüllt, um die Akzeptanz des Systems in Hinblick auf Leistungserwartung, Aufwand, sozialen Einfluss und Nutzungsvoraussetzungen zu erfassen.

4. SUS Fragebogen: Im Anschluss bewerteten die Teilnehmenden die Gebrauchstauglichkeit des jeweiligen Systems anhand der System Usability Scale

5. Post-Test (POT): Abschließend wurde ein Post-Test durchgeführt, um mögliche Veränderungen im Wissensstand, in der Einstellung oder emotionalen Reaktion durch die Interaktion zu messen.

Die Abfolge von UTAUT- und SUS-Fragebogen vor dem Post-Test wurde gewählt, um Eindrücke zur Akzeptanz und Gebrauchstauglichkeit direkt nach der Interaktion zu erfassen, bevor mögliche Reflexionseffekte im Post-Test einfließen konnten.

Für die gesamte Teilnahme am Experiment war ein Zeitfenster von etwa 20-25 Minuten vorgesehen. Diese strukturierte Abfolge gewährleistet eine fundierte Erhebung quantitativer und qualitativer Daten zur Nutzerakzeptanz und Interaktion mit KI-basierten Systemen.

3.1.2 Hypothesen

Die zentrale Forschungsfrage dieser Arbeit lautet: „Wie beeinflusst die Verwendung eines humanoiden Roboters im Vergleich zu einem Web-Chatbot mit RAG die Akzeptanz und die Benutzerfreundlichkeit im Bildungsbereich?“ werden im Folgenden die Hypothesen dieser Arbeit formuliert. Die Hypothesen konzentrieren sich auf die Unterschiede und mögliche Einflussfaktoren in der Akzeptanz sowie in der wahrgenommenen Benutzerfreundlichkeit zwischen der Nutzung eines humanoiden Roboters und eines webbasierten Chatbots mit RAG-Technologie im Bildungskontext.

3.1.2.1 Hypothesen zur System Usability Scale

Aufbauend auf der Forschungsfrage wurde eine konkrete Hypothese formuliert, die den Vergleich zwischen dem humanoiden Roboter NAO und dem webbasierten Chatbot fokussieren. Dabei steht die Dimensionen Benutzerfreundlichkeit im Mittelpunkt, die anhand der SUS-Scores [19] bewertet werden. Die folgenden Hypothesen bilden die Grundlage für die Überprüfung, ob sich Unterschiede in der wahrgenommenen Usability zwischen den beiden Systemen zeigen lassen.

H1: Der NAO-Roboter wird von den Nutzer*innen als benutzerfreundlicher wahrgenommen als der webbasierte Chatbot.

3.1.2.2 Hypothesen zur Technologieakzeptanz

Die Technologieakzeptanz wird im Rahmen dieser Arbeit anhand des UTAUT-Modells [18]untersucht, um Unterschiede in der Wahrnehmung und Nutzung zwischen dem NAO-Roboter und dem webbasierten Chatbot sichtbar zu machen. Im Mittelpunkt stehen dabei Faktoren wie Leistungserwartung, Aufwandserwartung, sozialer Einfluss, erleichternden Bedingungen sowie die Intention zur weiteren Nutzung. Auf Basis dieser Dimensionen wurden die folgenden Hypothesen formuliert, die den Vergleich der beiden Systeme systematisch überprüfen.

H2: Die Leistungserwartung ist beim NAO-Roboter höher ausgeprägt als beim Web-Chatbot.
H3: Die Aufwandserwartung ist beim Web-Chatbot höher ausgeprägt (d.h. leichter zu bedienen) als beim NAO-Roboter.
H4: Der Soziale Einfluss (wahrgenommene soziale Beeinflussung) ist beim NAO-Roboter stärker ausgeprägt als beim Web-Chatbot.
H5: Die erleichternden Bedingungen haben einen stärkeren Einfluss auf die Akzeptanz des NAO-Roboters als auf die des Web-Chatbots.
H6: Die Verhaltensabsicht zur weiteren Nutzung ist beim NAO-Roboter höher als beim Web-Chatbot.
H7: Der NAO erzielt höhere Akzeptanzwerte als der Web-Chatbot.

3.1.2.3 Hypothesen zur Wissensvermittlung und Lernverhalten

Ein zentraler Bestandteil der Untersuchung betrifft die Frage, inwiefern die beiden Systeme zur Wissensvermittlung beitragen und welches Lernverhalten die Nutzer*innen dabei zeigen. Um dies zu überprüfen, werden die Ergebnisse aus dem Pre- und Post-Test herangezogen, wodurch der tatsächliche Wissenszuwachs gemessen werden kann. Darüber hinaus wird aus der Literatur angenommen, dass der NAO-Roboter durch seine physische Präsenz und die damit verbundene soziale Interaktion einen stärkeren Lerneffekt erzielt als der Web-Chatbot. Schließlich wird auch der Zusammenhang zwischen der wahrgenommenen Usability und dem Wissenszuwachs betrachtet, da eine benutzerfreundliche Interaktion erfahrungsgemäß die Lernmotivation und den Lernerfolg positiv beeinflussen kann.

H8: Der Einsatz des NAO führt zu höheren Wissensvermittlung als der Einsatz eines webbasierten Chatbots.

3.1.3 Erfahrung und Allgemeine Fragen

Das Fragebogenkonstrukt (vgl. Tabelle 3) dient der Erfassung grundlegender demografischer Merkmale sowie der technologischen Vorerfahrung der Proband*innen. Ziel ist es, potenzielle Einflussfaktoren wie Alter, Geschlecht oder technologische Vertrautheit bei der späteren Datenanalyse berücksichtigen zu können.

Der Fragebogen wurde in ein eigenständiges Paket gegliedert, das vor dem Pre-Test durchgeführt wurde. Die enthaltenen Items Einwilligung (E), Allgemeine Fragen (AF) 1 bis 5 adressieren sowohl personenbezogene Merkmale als auch Erfahrungen mit digitalen Technologien.

Konstrukt	Frage
E	Bitte bestätigen Sie durch Ankreuzen beider Felder, dass Sie die Informationen verstanden haben und freiwillig teilnehmen möchten.
AF1	Welches System haben Sie verwendet?
AF2	Welches Geschlecht haben Sie?
AF3	Wie alt sind Sie?
AF4	Wie vertraut sind Sie allgemein mit KI-Systemen oder Chatbots (wie z. B. ChatGPT)?
AF5	Wie oft nutzen Sie digitale Lernmittel (z. B. Lernplattformen, Lernvideos, KI)?

Tabelle 3 - Allgemeine Fragen Konstrukt

3.1.4 Pretest und Posttest

Zur objektiven Messung von Wissensveränderungen vor und nach der Systeminteraktion wurde ein Pre- und Post-Test eingesetzt. Ziel dieses Tests ist es, fachbezogenes Wissen der Teilnehmenden im Themenbereich Künstliche Intelligenz, Datenverarbeitung und synthetische Daten zu erfassen und potenzielle Lerneffekte durch die Interaktion mit dem Web-Chatbot oder dem humanoiden Roboter zu identifizieren.

Der Test besteht aus fünf Fragen, die zentrale Konzepte aus den im Experiment behandelten Themenbereichen abdecken.

Die Fragenstruktur sowie die zugehörigen Antwortformate sind in Tabelle 4 dargestellt. Dabei steht PRT für die initiale Erhebung vor der Interaktion und POT für die Erhebung nach der Interaktion mit einem System. Die korrekten Antworten sind in Tabelle 5, Tabelle 6, Tabelle 7, Tabelle 8, und Tabelle 9 dargestellt.

Konstrukt	Frage
PRT1 / POT 1	Welche Aussage trifft auf Pseudonymisierung zu?
PRT2 / POT2	Welche der folgenden Maßnahmen wird nicht als klassische Bildmodifikation im Rahmen der Data Augmentation genannt?
PRT3 / POT3	Welcher der folgenden Begriffe beschreibt in der Datengüte-Definition die Abwesenheit von fehlenden Werten?
PRT4 / POT4	Wie kann man mit KI Bilder generieren?
PRT5 / POT5	Gib bitte drei Gründe warum man synthetische Daten braucht.

Tabelle 4 - Pre- und Posttest Fragen

Welche Aussage trifft auf Pseudonymisierung zu?

Antwortmöglichkeiten	**Richtig**
Persönliche Identifikatoren werden unwiederbringlich gelöscht	x
Die Beziehung zwischen Datensätzen in verschiedenen Tabellen bleibt erhalten.	✓
Es ist datenschutzrechtlich sicherer als vollständige Anonymisierung.	x
Eine Re-Identifikation (De-Pseudonymisierung) ist grundsätzlich ausgeschlossen.	x

Tabelle 5 - PRT1 / POT1 - richtige Antworten

Welche der folgenden Maßnahmen wird nicht als klassische Bildmodifikation im Rahmen der Data Augmentation genannt?

Antwortmöglichkeiten	**Richtig**
Rotieren vorhandener Bilder	✓
Spiegeln vorhandener Bilder	✓
Ändern der Helligkeit oder Farbe	✓
Segmentierung	x

Tabelle 6 - PRT2 / POT2 - richtige Antworten

Welcher der folgenden Begriffe beschreibt in der Datengüte-Definition die Abwesenheit von fehlenden Werten?

Antwortmöglichkeiten	**Richtig**
Correctness	✓
Completeness	✓
Consistency	✓
Uniqueness	✓
Uniformity	✓

Tabelle 7 - PRT3 / POT3 - richtige Antworten

Wie kann man mit KI Bilder generieren?

Antwortmöglichkeiten	**Richtig**
GAN	✓
Diffusion Modell	✓
Paint	x
LLM	x
VAE	✓
Cross Encoder	x
CNN	x
Auto Regression Models	✓
Log. Regression	x

Tabelle 8 - PRT4 / POT4 - richtige Antworten

Gib bitte drei Gründe warum man synthetische Daten braucht.

Antwortmöglichkeiten	**Richtig**
Anonymisierung von tabulare Daten	✓
Bekämpfung Bias & Ungleichheit der Klassen	✓
Statische Auswertung	x
Contentgenerierung für Marketing	x
Datenvisualiserung	x
Business Intelligence	x

Tabelle 9 - PRT5 / POT5 - richtige Antworten

3.1.5 UTAUT Fragebogen

Für die vorliegende Studie wurde der UTAUT-Fragebogen (vgl. Tabelle 10) anhand eines Originalfragebogens entwickelt. Die Struktur, Dimensionen und Itemformulierungen orientieren sich eng an der ursprünglichen Fassung, um die wissenschaftliche Fundierung und Vergleichbarkeit der Ergebnisse sicherzustellen. Gleichzeitig wurden die Formulierungen und Inhalte an das spezifisch untersuchte System angepasst, um eine möglichst realitätsnahe und kontextsensitive Erhebung zu gewährleisten.

Konstrukt	Konstrukt Nr.	Frage
Leistungserwartung	LE1	Ich empfinde das System in meinem Alltag als nützlich.
	LE2	Die Nutzung von dem System erhöht meine Chancen, Dinge zu erreichen, die mir wichtig sind.
	LE3	Die Nutzung von dem System hilft mir dabei, Dinge schneller zu erreichen.
	LE4	Die Nutzung von dem System erhöht meine Produktivität.
Aufwandserwartung	AE1	Mit dem System zu lernen ist einfach für mich.
	AE2	Die Interaktion mit dem System ist klar und verständlich.
	AE3	Ich finde es intuitiv, wie ich dem System Fragen stellen kann.
	AE4	Es ist unkompliziert, das System in mein Lernen einzubinden.
Sozialer Einfluss	SE1	Experten und Medien sind sich einig darin, Smart Mobility positiv zu bewerten.
	SE2	Menschen, die mein Verhalten beeinflussen, denken, dass ich das System nutzen sollte.
	SE3	Menschen, deren Meinung ich schätze, bevorzugen das System selbst.
	SE4	Ich habe das Gefühl, dass der Einsatz des System in meiner Lernumgebung akzeptiert ist.
erleichternde Bedingungen	EB1	Ich habe die notwendigen Ressourcen, um das System nutzen zu können
	EB2	Ich fühle mich ausreichend geschult, um das System produktiv zu nutzen.
	EB3	Die Lernumgebung unterstützt die Integration das System gut.
Verhaltensabsicht	VA1	Ich beabsichtige, den Chatbot regelmäßig für mein Lernen zu nutzen.

	VA2	Ich würde das System auch anderen Lernenden empfehlen.
	VA3	Ich werde das System bei zukünftigen Lernaufgaben einsetzen.

Tabelle 10 - UTAUT Fragebogen

3.1.6 System Usability Scale Fragebogen

Für die vorliegende Stichprobe wurde der System Usability Scale Fragebogen (vgl. Tabelle 11) in seiner ursprünglichen Struktur und Logik beibehalten und nur leicht an das untersuchte System angepasst. Dadurch bleibt die Vergleichbarkeit mit anderen Studien erhalten, gleichzeitig wird aber der Bezug zum spezifischen Anwendungskontext sichergestellt.

Konstrukt Nr.	**Frage**
SUS1	Ich würde dieses System häufig verwenden.
SUS2	Ich empfand das System als unnötig komplex.
SUS3	Ich fand das System einfach zu bedienen.
SUS4	Ich denke, dass ich die Unterstützung einer technisch versierten Person benötigen würde, um dieses System verwenden zu können.
SUS5	Ich fand, dass die verschiedenen Funktionen in diesem System gut integriert waren.
SUS6	Ich denke, dass es in diesem System zu viele Inkonsistenzen gibt.
SUS7	Ich kann mir vorstellen, dass die meisten Menschen lernen würden, dieses System sehr schnell zu benutzen.
SUS8	Ich empfand das System als umständlich in der Anwendung.
SUS9	Ich fühlte mich bei der Nutzung dieses Systems sehr sicher.
SUS10	Ich musste viele Dinge lernen, bevor ich mit diesem System effektiv arbeiten konnte.

Tabelle 11 - SUS Fragebogen

3.1.7 Likert Skala

Zur Erhebung der verschiedenen Konstrukte des UTAUT- und SUS-Modells wurden den Teilnehmenden standardisierte Aussagen vorgelegt, die anhand einer fünfstufigen Likert-Skala bewertet werden konnten (vgl. Tabelle 12) [18]. Diese Skala stellt ein etabliertes Instrument zur Messung subjektiver Einstellungen und Wahrnehmungen dar und erlaubt eine differenzierte Einschätzung der Zustimmung zu einzelnen Aussagen.

Die Skala reicht von 1 = „Stimme überhaupt nicht zu“ bis 5 = „Stimme voll und ganz zu“. Die Zwischenschritte ermöglichen es den Befragten, ihre Einschätzungen graduell zu differenzieren, wodurch sowohl starke als auch moderate Zustimmungs- oder Ablehnungstendenzen abgebildet werden können. Dieses Vorgehen gewährleistet eine hohe Vergleichbarkeit und Reliabilität der erfassten Daten im Hinblick auf die untersuchten Akzeptanzfaktoren.

Skala	**Information**
1	Stimme überhaupt nicht zu
2	Stimme nicht zu
3	Stimme weder zu noch nicht zu
4	Stimme zu
5	Stimme völlig zu

Tabelle 12 - Likert Skala

3.2 Versuchsgruppen

Zur Versuchsgruppe zählen Studierende der FH St. Pölten, die im Studiengang Data Science und AI im zweiten Semester immatrikuliert sind. Die Teilnehmer*innen wurden ausgewählt, um eine einheitliche Vergleichsbasis zu gewährleisten und Verzerrungen durch unterschiedliche Vorkenntnisse zu minimieren.

Da alle 20 Proband*innen aus demselben Studienbereich stammen, verfügen sie über einen ähnlichen Wissensstand im Bereich der Datenanalyse, Machine Learning (ML) und künstlichen Intelligenz. Durch diese fachliche Vertrautheit mit den zugrunde liegenden Technologien ist sichergestellt, dass sie die Funktionsweise des Systems besser verstehen und eine fundierte Bewertung der Benutzerfreundlichkeit (SUS) und Akzeptanz (UTAUT) abgeben können.

Die einheitliche Zusammensetzung der Gruppe ermöglicht eine klare Interpretation der Ergebnisse, da Unterschiede in der Nutzungserfahrung primär auf die Interaktionsform mit dem Web-Chatbot beziehungsweise dem NAO-Roboter zurückzuführen sind und nicht auf unterschiedliche Vorkenntnisse. Die Proband*innen dürfen entweder mit dem webbasierten Chatbot interagieren oder nur mit dem NAO Roboter.

3.3 Dokumente für die RAG Pipeline (Wissensdatenbank)

Die Wissensdatenbank-Dokumente stammen aus einem Kurs des Masterstudiengangs Data Intelligence der Fachhochschule St. Pölten und wurden den Studierenden vor dem Expierment nicht zur Verfügung gestellt. Die Dokumente bieten zunächst eine Einführung in Generative AI und erläutern die Unterschiede zwischen realen und synthetischen Daten. Anschließend wird der Begriff „synthetische Daten“ definiert und verschiedene Erzeugungsmethoden, von manueller Handarbeit über klassische Data-Augmentation bis hin zu KI-basierten Verfahren beschrieben. Die Folien stellen gängige Qualitätskriterien für synthetische Datensätze vor, etwa Korrektheit, Vollständigkeit und Konsistenz. Daran schließt sich eine Übersicht über statistische und ML-gestützte Methoden zur Messung der Ähnlichkeit zwischen Original- und synthetischen Daten an. Ein weiterer Schwerpunkt liegt auf den unterschiedlichen Generativmodellen wie Variational Autoencoders, GANs und Diffusionsmodellen sowie deren Vor- und Nachteilen. Die Studierenden erhalten zudem Einblicke in simulationsbasierte Ansätze zur Datenerzeugung und den Einsatz virtueller Umgebungen. Abschließend behandeln die Unterlagen praktische Anwendungsfälle und erläutern, wie synthetische Daten in Forschung und Industrie, etwa für Testsysteme oder Trainingsumgebungen, genutzt werden können. Dadurch bekommen die Studierenden einen umfassenden Einblick in Theorie, Methoden und Anwendungen synthetischer Daten.

3.4 Architektur

Die Architektur des Systems basiert auf einer RAG-Pipeline. Ziel ist es, externe Wissensquellen wie PDF‘s effizient zu nutzen und die Nutzungsinterkation über verschiedene Schnittstellen zu ermöglichen.

Das System umfasst mehrere Kernkomponenten, die in einem modularen Aufbau miteinander interagieren. Der gesamte Ablauf kann in zwei Hauptphasen unterteilt werden, zum einen Dokumente ingesten und eine Query absetzen.

3.4.1 Dokument Ingestion

Der Dokumenten-Ingest-Prozess ist ein zentraler Bestandteil der Architektur und ermöglicht die Verarbeitung, Strukturierung und Indexierung von unstrukturierten Dokumenten, um sie für eine effiziente Abfrage mittels RAG nutzbar zu machen. Dabei wird LlamaIndex in Kombination mit einem Natural Language Model (NLM) Ingestor und einem PostgreSQL-basierten Vektorstore eingesetzt.

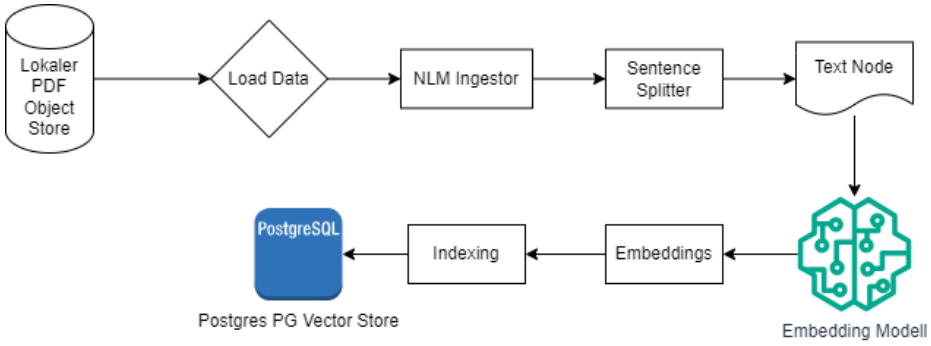

Abbildung 3 - Ingest Prozess

3.4.1.1 Prozessübersicht

Der Ingest-Prozess folgt mehreren aufeinander abgestimmten Schritten, die in der Abbildung dargestellt sind. Zunächst werden die Dokumente aus einer lokalen PDF-Datenbank oder einem Objektspeicher geladen. Neben PDF-Dateien können auch alternative Formate wie TXT oder DOCX verarbeitet werden.

Anschließend erfolgt der Datenladeprozess, in dem die Rohdaten in das System importiert und für die weitere Verarbeitung vorbereitet werden. In dieser Phase können Metadaten extrahiert werden, um zusätzliche Informationen über die Dokumente bereitzustellen.

Die eigentliche Verarbeitung übernimmt der NLM Ingestor, der eine semantische Vorverarbeitung der Texte vornimmt. Dabei kommen Natural Language Processing (NLP)-Techniken wie Optical Character Recognition (OCR), Textextraktion und Normalisierung zum Einsatz, um die Dokumente in ein strukturiertes Format zu extrahieren.

Nach der Vorverarbeitung erfolgt die Segmentierung der Texte durch den Sentence Splitter. In diesem Schritt werden die Dokumente in kleinere Textabschnitte unterteilt, um eine granulare Suchbarkeit zu ermöglichen. Die Segmentierung kann satzweise, absatzweise oder anhand semantischer Cluster erfolgen, um eine sinnvolle Struktur für spätere Abfragen zu schaffen.

Die so gewonnenen Chunks werden anschließend als Text Nodes gespeichert, die als zentrale Verarbeitungseinheiten für die semantische Suche dienen. Um die inhaltliche Bedeutung der Texte maschinell erfassbar zu machen, wird ein Embedding-Modell von OpenAI Namens „text-embedding-3-small“ verwendet. Dieses Modell wandelt die Textknoten in hochdimensionale Vektoren um, die die semantische Bedeutung der Inhalte repräsentieren. Dadurch wird eine ähnlichkeitsbasierte Suchfunktion ermöglicht, bei der verwandte Konzepte unabhängig von ihrer exakten Wortwahl erkannt werden können.

Die Embeddings werden abschließend in einer PostgreSQL PG Vector Store-Datenbank gespeichert, die speziell für Vektorbasierte Suchen optimiert ist. Diese Speicherung erlaubt eine effiziente semantische Suche nach thematisch relevanten Dokumenten und ermöglicht eine schnelle Wissensabfrage für RAG-Systeme.

3.4.1.2 Bedeutung für das System

Dieser Prozess ermöglicht eine strukturierte Speicherung sowie eine schnelle Abfrage großer Mengen an unstrukturierten Daten. Durch die Kombination von LlamaIndex, dem NLM Ingestor und einem Embedding-Modell wird sichergestellt, dass relevante Inhalte effizient durchsucht und für KI-gestützte Antworten bereitgestellt werden können. Besonders hervorzuheben sind dabei drei zentrale Vorteile. Zum einen bietet der Ansatz eine hohe Skalierbarkeit, sodass auch große Mengen an Textdaten in Echtzeit verarbeitet werden können. Zum anderen ermöglicht die Nutzung von Embeddings eine präzise semantische Suche, bei der relevante Inhalte auch ohne exakte Wortübereinstimmung zuverlässig identifiziert werden. Darüber hinaus sorgt die Einbindung des PostgreSQL PG Vector Store für eine effiziente Speicherung, die schnelle Ähnlichkeitsvergleiche und damit eine performante Abfrage der Daten unterstützt.

3.4.2 Query Komponente

Die Query-Komponente bildet das Herzstück der Interaktionslogik des Systems und verarbeitet Benutzeranfragen, um relevante Antworten aus dem RAG-System zu generieren. Diese Komponente ist entscheidend für die Verbindung zwischen dem Frontend (Web-Chatbot oder NAO-Roboter) und dem Backend, dass die semantische Suche, die Informationswiedergewinnung und die Antwortgenerierung übernimmt. Unabhängig davon, ob der oder die Nutzer*in über den Web-Chatbot oder den humanoiden NAO-Roboter interagiert, bleibt das Backend unverändert und folgt einem standardisierten Ablauf zur Beantwortung von Anfragen.

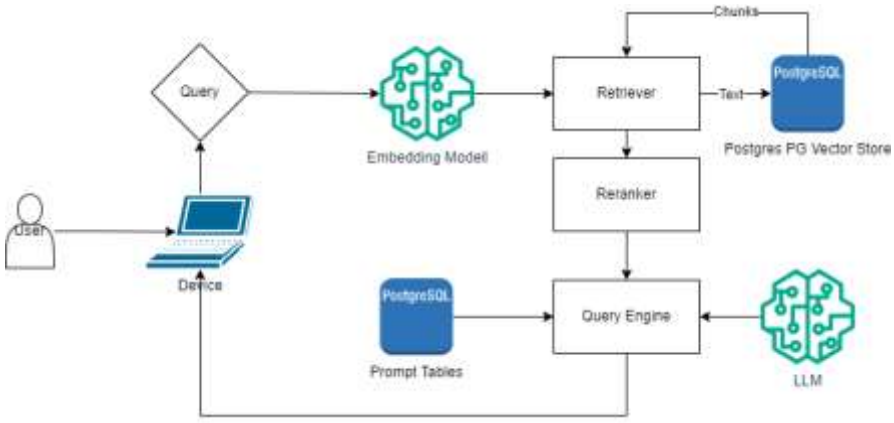

Abbildung 4 - Query Prozess

3.4.2.1 Pozess der Query-Verarbeitung

Die Verarbeitung einer Benutzeranfrage (vgl. Abbildung 4) erfolgt in mehreren aufeinanderfolgenden Schritten, um eine präzise und kontextbezogene Antwort zu generieren. Der Prozess beginnt mit der Benutzereingabe und Anfrageübermittlung. Die Nutzer*innen können entweder über den Webchatbot oder den NAO-Roboter eine Text- oder Spracheingabe tätigen. Falls die Eingabe in gesprochener Form erfolgt, wird sie durch eine Speech-to-Text-Technologie in Text umgewandelt. Anschließend wird die Anfrage an das Backend weitergeleitet, wo sie zunächst in eine Vektorrepräsentation überführt wird.

Im nächsten Schritt erfolgt die Erstellung von Embeddings, bei der das Embedding-Model die Anfrage in einen hochdimensionalen Vektor umwandelt. Diese Vektordarstellung ermöglicht es, in der Datenbank nach inhaltlich ähnlichen Textabschnitten zu suchen, um relevante Informationen für die spätere Antwort zu extrahieren.

Nach der Vektorisierung übernimmt der Retriever die Aufgabe, die Benutzeranfrage mit den zuvor im PostgreSQL PG Vector Store gespeicherten Dokumenten-Embeddings von dem Modell „text-embedding-3-small“ zu vergleichen. In diesem Schritt werden die relevantesten Dokumenten-Chunks identifiziert und zur weiteren Verarbeitung weitergeleitet. Da die erste Auswahl oft eine größere Menge potenziell relevanter Texte zurückliefert, wird eine Neugewichtung der Ergebnisse durch den Reranker durchgeführt. Dieser priorisiert die relevantesten Chunks und verbessert dadurch die Qualität der abgerufenen Informationen, sodass die bestmögliche Antwort generiert werden kann.

Um die Nutzeranfrage weiter zu optimieren, erfolgt eine Erweiterung mit Informationen aus den Prompt Tables, die in einer separaten PostgreSQL-Datenbank gespeichert sind. Diese Tabellen enthalten vordefinierte Systemprompts, Benutzerhistorien oder spezielle Formatierungsanweisungen, die die Antwortqualität des Sprachmodells verbessern.

Schließlich werden die verarbeiteten Daten an das Large Language Model (LLM) weitergeleitet. Das LLM kombiniert die ursprüngliche Nutzeranfrage mit den abgerufenen Dokumenten und generiert eine maßgeschneiderte, kontextbezogene Antwort, die sowohl die aktuellen Eingaben als auch das zuvor indexierte Wissen berücksichtigt.

Die fertige Antwort wird anschließend an das Frontend zurückgesendet. Im Fall des Web-Chatbots erfolgt die Darstellung als klassische Textnachricht, während beim NAO-Roboter zusätzlich eine Text-to-Speech-Engine zum Einsatz kommt. Dadurch kann der Roboter die generierte Antwort nicht nur anzeigen, sondern auch verbal kommunizieren, um eine noch intuitivere Mensch-Roboter-Interaktion (MRI) zu ermöglichen.

3.4.2.2 Bedeutung und Vorteile der Architektur

Die Query-Komponente bietet mehrere entscheidende Vorteile, die zur Leistungsfähigkeit des Gesamtsystems beitragen. Durch den Retriever-Reranker-Ansatz können relevante Inhalte schnell und präzise identifiziert werden, was eine besonders effiziente Informationsverarbeitung ermöglicht. Gleichzeitig sorgt die Kombination aus Embedding-basiertem Retrieval und Prompt-Optimierung für eine erweiterte Kontextualisierung, wodurch die Qualität der Antworten eines LLM deutlich verbessert wird. Ein weiterer Vorteil liegt in der Plattformunabhängigkeit, da derselbe Backend-Prozess sowohl für den webbasierten Chatbot als auch für den NAO-Roboter genutzt werden kann, was eine flexible und vielseitige Bereitstellung erlaubt. Zudem unterstützt die Komponente eine semantische Suche in unstrukturierten Dokumenten, indem ein PostgreSQL PG Vector Store eingesetzt wird, der die Speicherung und schnelle Abfrage von Textsegmenten ermöglicht, die mit klassischen Keyword-Suchen nicht effizient gefunden werden könnten.

3.4.2.3 Experiment NAO

Die Integration des NAO-Roboters in das bestehende System stellt eine besondere Herausforderung dar, da er auf der veralteten Python-2.7-Umgebung basiert.

Das Ziel dieses Experiments ist es, die Sprachverarbeitung, Steuerung und Antwortgenerierung des NAO-Roboters innerhalb einer modularen Container-Architektur zu realisieren. Dabei werden RAG und Speech Recognition betrieben, während die direkte Interaktion mit dem NAO-Roboter über einen Python-2.7-Container mit NAOqi erfolgt. Das RASA CAI und RAG Framework wird in einer Docker Umgebung initalisiert.

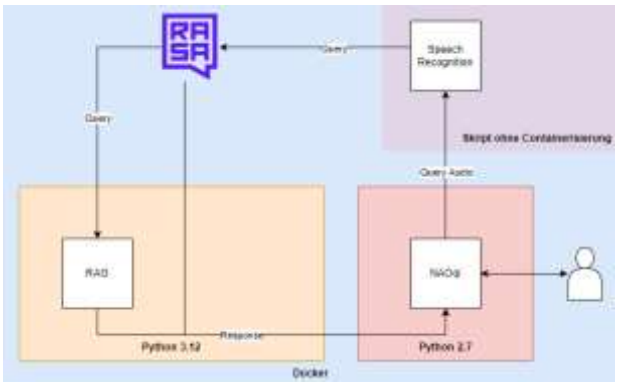

Abbildung 5 - NAO Containerisierung (Docker)

Die containerisierte Architektur ermöglicht eine flexible, skalierbare und interoperable Kommunikation zwischen den Software-Modulen. Der Ablauf der Interaktion zwischen dem NAO-Roboter und der RAG-basierten KI erfolgt in mehreren Schritten.

Zunächst stellen die Nutzer*innen dem NAO-Roboter eine Frage oder gibt einen Sprachbefehl. Die Audioeingabe wird im NAOqi-Container (Python 2.7) verarbeitet und anschließend an die Speech Recognition-Komponente in einen lokalen Python-3.12 Skript weitergeleitet. Dort wird die Spracheingabe durch ein Speech Recognition-Modell in Text umgewandelt. Das musste leider so umgesetzt werden, da einige Komponenten nicht im Docker-Container ausgeführt werden können. Die transkribierte Anfrage wird daraufhin an das RASA-Modell weitergeleitet, das die Intentions leitet oder eine Response direkt an den NAO wieder weitergibt.

Sobald die relevanten Inhalte extrahiert wurden, erfolgt die Antwortgenerierung durch das RAG-System. Dieses durchsucht die Wissensdatenbank nach passenden Informationen und übergibt die aufbereitete Antwort an den FastAPI Endpoint.

Die generierte Antwort wird anschließend zurück an den NAOqi-Container gesendet. Dort wird sie in Sprache umgewandelt, sodass der NAO-Roboter die Antwort verbal ausgeben kann. Ergänzend zur Sprachwiedergabe können gestische oder visuelle Interaktionen erfolgen, um die Kommunikation mit dem Nutzer intuitiver und natürlicher zu gestalten.

3.4.2.4 Experiment Web Chatbot

Durch den Web-Chatbot wird eine benutzerfreundliche Schnittstelle bereitgestellt, über die Nutzer*innen mit dem System über eine Weboberfläche kommunizieren können. Die Architektur basiert auf Python 3.12 und ermöglicht eine Interaktion zwischen den verschiedenen Komponenten.

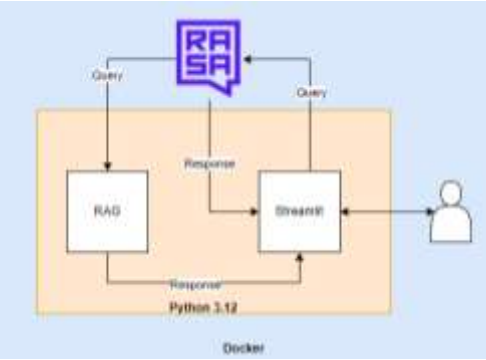

Abbildung 6 – Web-Chatbot Containerisierung

Die Interaktion zwischen dem Anwender*innen und dem Chatbot erfolgt über mehrere aufeinanderfolgende Verarbeitungsschritte.

Der Prozess beginnt mit der Eingabe einer Anfrage durch die Anwender*innen in das Streamlit-Frontend. Diese Anfrage wird an das RAG-Modell weitergeleitet, das die semantische Suche in einer Vektor-Datenbank durchführt. Parallel dazu wird die Anfrage an RASA übermittelt, das für die Dialogsteuerung und Kontextverwaltung zuständig ist.

Sobald RAG relevante Informationen aus der Wissensbasis extrahiert hat, wird die Antwort generiert und an Streamlit zurückgesendet. Das Frontend zeigt die Antwort in einer übersichtlichen und verständlichen Form für die Nutzer*innen an. Falls erforderlich, können die Anwender*innen Rückfragen stellen oder den Dialog fortsetzen, wobei RASA die Konversation verwaltet und sicherstellt, dass der Kontext erhalten bleibt.

3.5 Eingesetzte Technologien

In diesem Kapitel werden die zentralen Technologien vorgestellt, die für die Umsetzung der Arbeit genutzt wurden. Die Auswahl der eingesetzten Frameworks und Tools basiert auf ihren spezifischen Stärken in Bezug auf Datenverarbeitung, LLMs, Embedding Modelle, Benutzerinteraktion und Robotik.

3.5.1 Python Bibliotheken

Python ist eine der am weitesten verbreiteten Programmiersprachen in der künstlichen Intelligenz, Datenverarbeitung und Softwareentwicklung. Ihre Beliebtheit verdankt die Sprache nicht nur ihrer klaren und einsteigerfreundlichen Syntax, sondern auch der umfangreichen Sammlung an Bibliotheken und Frameworks, die vielfältige Anwendungsbereiche abdecken, von numerischer Berechnung (z. B. NumPy, pandas) über maschinelles Lernen (z. B. scikit-learn, TensorFlow) bis hin zur Webentwicklung (z. B. Flask, Django)

Im Rahmen dieses Experiments kamen sowohl Python 3.12 als auch Python 2.7 zum Einsatz. Der Einsatz beider Versionen war notwendig, um Kompatibilität mit bestehenden Abhängigkeiten und Modulen sicherzustellen, die teilweise noch nicht auf Python 3 portiert wurden. Während Python 3 als aktueller Standard verwendet wurde, diente Python 2.7 insbesondere dem Zugriff auf ältere Bibliotheken wie NAOqi, die in der aktuellen Version nicht mehr unterstützt werden.

Für die Umsetzung des Experiments wurden diverse Python-Bibliotheken verwendet, um Daten zu verarbeiten, Modelle zu integrieren sowie Visualisierungen und Schnittstellen bereitzustellen.

3.5.1.1 LlamaIndex

Das LlamaIndex-Framework dient der Verarbeitung, Strukturierung und dem Zugriff auf private oder domänenspezifische Daten in LLM-Anwendungen. Während Chat Completion primär auf die Integration von Dialogverläufen ausgerichtet ist, liegt der Fokus von LlamaIndex auf der effizienten Suche und Einbindung externer Informationen. Dabei nutzt es moderne RAG-Techniken, um die Qualität der Modellantworten zu verbessern. Auf diese Weise kann ein LLM über sein vortrainiertes Wissen hinaus auf aktuelle oder organisationsspezifische Daten zugreifen. [211]

Ein zentrales Merkmal von LlamaIndex ist seine flexible Schnittstelle, die eine nahtlose Integration in bestehende KI-Systeme ermöglicht. Zu den wesentlichen Komponenten gehören unter anderem die Datenquellen-Handler, die die Verarbeitung verschiedener Formate unterstützen. Der vektorbasierte Index erstellt Embeddings aus den Daten und speichert diese in einer durchsuchbaren Struktur, wobei unterschiedliche Vektor Stores wie Qdrant oder Postgres verwendet werden können. Über den Retriever werden anschließend relevante Daten aus dem Index basierend auf Nutzeranfragen ermittelt. Die Query-Engine verbindet die gefundenen Informationen mit einem LLM und generiert daraus präzise und kontextreiche Antworten. [211]

Insgesamt stellt LlamaIndex ein leistungsfähiges Tool für die effiziente Nutzung externer Informationen in LLM-Anwendungen dar. Durch die Kombination von Vektor-Datenbanken, Retrieval-Methoden und großen Sprachmodellen ermöglicht es eine deutliche Verbesserung der Antwortqualität und Kontextgenauigkeit. Gleichzeitig erlaubt die modulare Architektur eine einfache Anpassung an unterschiedliche Anwendungsfälle, was LlamaIndex besonders vielseitig einsetzbar macht. [211]

3.5.1.2 Streamlit

Die Entwicklung von benutzerfreundlichen und interaktiven Anwendungen für KI-gestützte Systeme stellt eine zentrale Herausforderung dar. Häufig erfordert dies eine Kombination aus Webentwicklung, Datenverarbeitung und Visualisierung. Streamlit ist ein Open-Source-Framework (OSF), das speziell für die einfache und schnelle Erstellung von datengetriebenen Webanwendungen entwickelt wurde. Es ermöglicht die direkte Integration von Machine Learning-Modellen, Datenvisualisierungen und interaktiven Benutzeroberflächen mit minimalem Programmieraufwand. [212]

Streamlit basiert auf Python und erlaubt es Entwicklern, mit nur wenigen Codezeilen interaktive Anwendungen zu erstellen. Die Architektur folgt einem deklarativen Ansatz, wodurch die Benutzeroberfläche automatisch aktualisiert wird, sobald sich die zugrunde liegenden Daten ändern. [212]

Durch den Einsatz von Streamlit wird eine einfache, aber leistungsstarke Schnittstelle für die KI-Anwendung geschaffen. In Kombination mit LlamaIndex ermöglicht es die Erstellung einer interaktiven Plattform, die den Abruf und die Verarbeitung von externem Wissen für große Sprachmodelle (LLMs) optimiert. Die Benutzerfreundlichkeit und die schnelle Entwicklung machen Streamlit zu einer idealen Wahl für dieses Experiment als Web-Chatbot. [212]

3.5.1.3 NAOqi

NAOqi ist das proprietäre Middleware-Framework von SoftBank Robotics zur Steuerung humanoider Roboter wie NAO und Pepper. Es stellt eine modulare Architektur bereit, die es ermöglicht, verschiedene funktionale Komponenten wie Bewegung, Sprachverarbeitung, Sensorik oder Kamerazugriff zu koordinieren und über standardisierte Schnittstellen zu programmieren. [213]

Die Interaktion mit NAOqi erfolgt über eine API, die in mehreren Programmiersprachen zur Verfügung steht, insbesondere in Python und C++. Für Python-Entwicklungen stellt SoftBank Robotics ein spezifisches SDK bereit, welches die Kommunikation mit den internen Modulen des Roboters ermöglicht, z. B. über das ALProxy-Konzept. [213]

Ein wesentlicher technischer Aspekt ist, dass NAOqi in seiner aktuellen Version ausschließlich mit Python 2.7 kompatibel ist. Diese Abhängigkeit ergibt sich aus der ursprünglichen Entwicklungszeit des Frameworks, dass viele der eingebundenen nativen Module auf der Laufzeitumgebung von Python 2.7 basieren. Eine Ausführung unter Python 3.x ist daher nicht möglich, ohne tiefgreifende Anpassungen oder Re-Implementierungen vorzunehmen. [213]

Diese Einschränkung stellt eine Herausforderung für moderne Softwareprojekte dar, da Python 2.7 seit dem 1. Januar 2020 offiziell das Ende des Supports erreicht hat. Dennoch ist der Einsatz von Python 2.7 im Kontext von NAOqi derzeit alternativlos, sofern keine Migration auf andere Frameworks oder Hardwareplattformen erfolgt. [213]

3.5.2 PostgresSQL

PostgreSQL ist ein objektrelationales Datenbankmanagementsystem (ORDBMS), das seit den 1990er Jahren kontinuierlich weiterentwickelt wird und sich durch hohe Stabilität, Erweiterbarkeit sowie vollständige Unterstützung des SQL-Standards auszeichnet. Es ist unter einer Open-Source-Lizenz verfügbar und eignet sich aufgrund seiner ACID-Konformität besonders für Anwendungen mit hohen Anforderungen an Datenintegrität und Transaktionssicherheit. [214]

In der vorliegenden Arbeit wurde PostgreSQL als zentrale persistente Datenhaltungskomponente eingesetzt. Zur Speicherung und Verarbeitung von Vektorrepräsentationen, insbesondere die generierte Embeddings, kam die PostgreSQL-Erweiterung pgvector zum Einsatz. [214]

Die Integration von pgvector erlaubt semantische Suchen direkt in der Datenbank, ohne externe Vektorsuchsysteme wie FAISS oder Pinecone einbinden zu müssen. Dies bietet erhebliche Vorteile hinsichtlich Performance, Wartbarkeit und Datenschutz, insbesondere bei lokalen oder selbst gehosteten Anwendungen. Die Vektorsuche innerhalb von PostgreSQL kann durch Indexierung mit Approximate Nearest Neighbor (ANN) weiter beschleunigt werden, wodurch auch größere Embedding-Mengen effizient verarbeitet werden können. [214]

Die Kombination aus PostgreSQL und pgvector stellt somit eine moderne, performante und erweiterbare Lösung für datenbankgestützte KI-Anwendungen dar, bei denen sowohl strukturierte Informationen als auch semantische Vektordaten gemeinsam gespeichert und abgefragt werden müssen.[214]

3.5.3 RASA

RASA ist eine leistungsstarke Open-Source-Plattform für Conversational AI (CAI), die speziell für die Entwicklung von chatbot- und sprachgesteuerten Assistenten konzipiert wurde. Sie ermöglicht die Erstellung intelligenter und kontextbewusster Dialogsysteme, die flexibel in verschiedene Anwendungen integriert werden können.[215]

Das Framework bietet eine Kombination aus Natural Language Understanding (NLU) zur Erkennung von Benutzerabsichten und Dialogue Management, um dynamische Gespräche zu steuern. Dadurch eignet sich RASA besonders für maßgeschneiderte KI-Assistenten, die über reine Regelwerke hinausgehen und mit maschinellem Lernen verbessert werden können.

Im Experiment werden zwei Intents unterschieden (vgl. Tabelle 13). Der Intent „Chatbot_intention_question_mode“ versetzt den Assistenten in einen Fragemodus, in dem er gezielt Rückfragen oder Quiz- und Lernfragen stellt. Der andere Intent „chatbot_intention_answer_mode“ aktiviert einen Antwortmodus, in dem er Nutzerfragen erklärt und beantwortet. Die Erkennung erfolgt über Rasa NLU anhand von Beispieläußerungen. Nach der Zuordnung wird ein System-Prompt gesetzt, der die Rolle des Assistenten klar festlegt. Eine Python-Logik übernimmt anschließend die Umsetzung. Im Fragemodus erzeugt sie passende Fragen, im Antwortmodus greift sie auf die Wissensbasis oder das Sprachmodell zu und liefert eine präzise Antwort. Auf diese Weise bleibt der Backend-Prozess einheitlich, während der Chatbot je nach Nutzerabsicht flexibel zwischen Fragen und Antworten wechselt.

Modus	Intent	Zweck	Intent‘s (Auszug)
Fragemodus	Chatbot intention question mode	Nutzer*in möchte, dass der Bot Fragen stellt	„Stelle mir Fragen“, „Frag mich etwas“, „Ich möchte Fragen beantworten“, „Gib mir bitte eine Frage“
Antwortmodus	Chatbot intention answer mode	Nutzer*in stellt eine Frage an den Bot	„Was bedeutet RAG?“, „Wie funktioniert Rasa?“, „Was ist CRISP DM?“

Tabelle 13 - Modus RASA und RAG

3.5.4 Docker

Docker ist eine leistungsstarke Containerisierungsplattform, die es ermöglicht, Anwendungen und deren Abhängigkeiten in isolierten Containern bereitzustellen. Durch den Einsatz von Docker wird sichergestellt, dass alle Systemkomponenten unabhängig von der zugrunde liegenden Infrastruktur konsistent ausgeführt werden können. Dies verbessert die Portabilität, Skalierbarkeit und Wartbarkeit der Anwendung und ermöglicht eine effiziente Bereitstellung sowohl in lokalen Entwicklungsumgebungen als auch in der Cloud. [216]

In dieser Arbeit wird Docker verwendet, um die verschiedenen Module des RAG-Systems, darunter LlamaIndex, FastAPI, Streamlit, PostgreSQL PG Vector Store und den NAO-Roboter, in containerisierten Umgebungen auszuführen. Dies gewährleistet eine einheitliche und reproduzierbare Umgebung für alle Systemkomponenten.

3.5.5 NLM Ingestor

Der NLM Ingestor ist eine weitere zentrale Komponente des Systems, die für die Vorverarbeitung und Strukturierung von Dokumenten verantwortlich ist. Er übernimmt die Aufgabe, unstrukturierte Daten aus verschiedenen Quellen (z. B. PDFs oder Textdateien) zu analysieren, in kleinere Segmente zu unterteilen und für die weitere Verarbeitung durch LlamaIndex vorzubereiten. [211], [217]

Da der Ingestor als eigenständiger Service innerhalb der Architektur arbeitet, wird er in einem Docker-Container bereitgestellt. Durch die Containerisierung wird eine konsistente und reproduzierbare Umgebung geschaffen, die unabhängig von der zugrunde liegenden Infrastruktur funktioniert. [217]

3.5.6 Modelle

In diesem Kapitel werden die zentralen Modelle beschrieben, die für die Umsetzung des Experiments verwendet wurden. Embedding Modelle und Große Sprachmodelle (LLMs) benötigen verschiedene Arten von Repräsentationen und Algorithmen, um effizient mit externen Datenquellen zu interagieren.

3.5.6.1 Embedding Model

Embedding-Modelle (EM) spielen eine zentrale Rolle bei der Verarbeitung natürlicher Sprache, insbesondere in Systemen, die semantische Suche, Textklassifikation oder kontextuelle Ähnlichkeitsmessungen erfordern. Sie wandeln Text in hochdimensionale Vektoren um, die es ermöglichen, Bedeutungsähnlichkeiten zwischen Begriffen, Sätzen oder Dokumenten effizient zu berechnen. [211]

Für das Experiment das „text-embedding-3-small“ Embedding-Modell von OpenAI eingesetzt. Die Entscheidung für dieses Modell fiel aus mehreren Gründen. Zum einen ermöglicht die einfache Integration über die OpenAI API eine unkomplizierte Nutzung, da es sich um einen cloudbasierten Dienst handelt und somit keine lokale Modellbereitstellung erforderlich ist. Dies erleichtert die Implementierung erheblich. Darüber hinaus überzeugt das Modell durch seine hohe Performance und Genauigkeit, da „text-embedding-3-small““ im Vergleich zu früheren Versionen verbesserte semantische Repräsentationen von Texten liefert und dadurch eine präzisere semantische Suche ermöglicht. Ein weiterer Vorteil ist die Skalierbarkeit, da die Berechnung der Embeddings vollständig in der Cloud erfolgt, lassen sich auch große Mengen an Textdaten effizient verarbeiten, ohne dass dafür leistungsstarke lokale Hardware notwendig wäre.

3.5.6.2 Foundation Model

Im Experiment wird das GPT-4o-Modell von OpenAI als Foundation LL-Model eingesetzt. Die Entscheidung für dieses Modell beruht auf mehreren zentralen Faktoren. Ein wesentlicher Vorteil ist die einfache Integration über die OpenAI API, da GPT-4o direkt über eine REST-API angesprochen werden kann. Dadurch lässt sich das Modell unkompliziert in bestehende Anwendungen einbinden, was den Implementierungsaufwand im Vergleich zu lokal gehosteten Modellen erheblich reduziert. Hinzu kommt die hohe Leistungsfähigkeit, GPT-4o verfügt über eine optimierte Architektur, die verbesserte Antwortzeiten und eine gesteigerte Effizienz im Vergleich zu früheren Versionen ermöglicht. Als Foundation Model bringt es zudem ein aktuelles und breites vortrainiertes Wissen mit, das für vielseitige Anwendungsfälle genutzt werden kann. In Kombination mit LlamaIndex besteht darüber hinaus die Möglichkeit, dieses Wissen gezielt, um domänenspezifische Informationen zu erweitern. Ein weiterer Vorteil liegt in der Skalierbarkeit und Wartung, da das Modell vollständig von OpenAI gehostet wird. Damit entfällt die Notwendigkeit, eigene Rechenressourcen für Training oder Wartung bereitzustellen, was bei begrenzter Infrastruktur die Nutzung leistungsfähiger Modelle ermöglicht.

3.5.7 NAO Roboter

Es wird der NAO-Roboter als physische Schnittstelle für die Interaktion mit einem LLM-gestützten Assistenzsystem eingesetzt. Ziel ist es, die Wahrnehmung, Nutzbarkeit und Akzeptanz einer humanoiden KI in unterschiedlichen Interaktionsszenarien zu evaluieren. Der Roboter übernimmt dabei die Rolle eines sprachgesteuerten Assistenten, der mit den Nutzer*innen kommuniziert. Neben der sprachbasierten Interaktion, bei der Benutzereingaben verarbeitet und Antworten mithilfe der RAG-Komponente generiert werden, spielt auch die nonverbale Kommunikation eine wichtige Rolle. Durch Kopfbewegungen, Gesten und LED-Anzeigen unterstützt NAO die natürliche Interaktion und trägt dazu bei, die Immersion im Dialog zu erhöhen.

3.5.8 Interaktionstechnologie

Für die Kommunikation zwischen den verschiedenen Komponenten der Anwendung wird FastAPI eingesetzt. Dabei handelt es sich um ein leistungsstarkes und modernes Web-Framework für die Entwicklung von RESTful APIs in Python, das sich insbesondere durch seine hohe Geschwindigkeit, einfache Handhabung und die eingebaute Unterstützung für asynchrone Verarbeitung auszeichnet. Ein wesentlicher Vorteil von FastAPI liegt in seiner Performance, da es auf Starlette und Pydantic basiert und damit zu den schnellsten Python-APIs zählt. Darüber hinaus bietet es eine automatische Generierung von API-Dokumentationen, da dank der OpenAPI- und Swagger-Integration sämtliche Endpunkte unmittelbar dokumentiert werden. Durch die native Unterstützung von asynchroner Verarbeitung mittels async/await können Anfragen effizient und skalierbar verarbeitet werden, was die Anwendung besonders leistungsfähig macht. Zudem ermöglicht FastAPI eine einfache Integration mit LLMs und Datenbanken, sodass beispielsweise LlamaIndex, Embedding-Modelle oder externe Datenquellen problemlos angebunden werden können. Diese Eigenschaften machen FastAPI zu einer idealen Wahl für die Umsetzung der Kommunikationslogik in dieser Anwendung. [218]

In diesem Experiment-Setting fungiert FastAPI als zentrale Schnittstelle für die Interaktion zwischen Benutzer, der RAG-Pipeline und externen Datenquellen. Über die API werden Nutzereingaben, beispielsweise aus Streamlit oder vom NAO-Roboter, verarbeitet und an die RAG-Pipeline weitergeleitet. Gleichzeitig stellt FastAPI eine Web-Schnittstelle bereit, über die sowohl Streamlit als auch andere Clients auf das System zugreifen können. Durch den Einsatz von FastAPI wird gewährleistet, dass die Kommunikation zwischen den Modellen und der Benutzeroberfläche effizient, skalierbar und flexibel erweiterbar bleibt.

3.6 Methoden der statistischen Analyse

Zur systematischen Auswertung der erhobenen Daten wurden verschiedene statistische Analysemethoden herangezogen, die es ermöglichen, zentrale Tendenzen, Streuungen sowie die Qualität der eingesetzten Messinstrumente zu bewerten. Dabei steht die deskriptive Statistik im Vordergrund, die einen Überblick über grundlegende Kennzahlen wie Mittelwert, Minimal- und Maximalwerte sowie Standardabweichung liefert. Ergänzend werden Maße zur Verteilungsform der Daten, Schiefe und Kurtosis, berücksichtigt, um mögliche Abweichungen

von der Normalverteilung sichtbar zu machen. Darüber hinaus wird die interne Konsistenz der eingesetzten Skalen mithilfe von Cronbach's Alpha überprüft.

3.6.1 Deskriptive Statistik

Um die Forschungsfrage beantworten zu können, wurde eine deskriptive Statistik durchgeführt. Die erhobenen Daten wurden dabei mithilfe statistischer Methoden wie dem Minimalwert, dem Mittelwert, dem Maximalwert sowie der Standardabweichung ausgewertet. Die deskriptive Statistik dient dazu, die erhobenen Werte zusammenzufassen, zu beschreiben und charakteristische Merkmale, Muster oder Trends in den Daten sichtbar zu machen.

Der Minimalwert (min) einer Datenreihe wird durch den Vergleich sämtlicher Werte ermittelt, wobei der kleinste Wert das Ergebnis darstellt. Im Gegensatz dazu beschreibt der Maximalwert (max) den größten Wert einer Datenreihe und wird auf dieselbe Weise bestimmt

Der Mittelwert, auch als arithmetisches Mittel oder Durchschnittswert ($\bar{x}$) bezeichnet, wurde in dieser Arbeit zur Bestimmung der zentralen Tendenz verwendet. Für seine Berechnung wurden die Werte der einzelnen Teilnehmenden innerhalb eines Fragenpakets sowie in den jeweiligen Gruppen (z. B. Geschlecht, Erfahrung, Bildungsniveau, System) summiert und durch die Anzahl der Teilnehmenden geteilt. Damit lassen sich die Fragen und Fragenpakete des erweiterten UTAUT-Modells nach Heerink et al. [219] messbar machen. Wie in der Literatur empfohlen, sollten Mittelwerte für jede Frage und jedes Konstrukt gebildet werden, um eine valide Auswertung zu gewährleisten. [219]

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} a_k$$

Formel 1 - arithmetisches Mittel

Ein weiteres Maß ist die Standardabweichung (σ), die angibt, wie stark die einzelnen Werte um den Mittelwert variieren. Sie dient zur Quantifizierung der Streuung und damit zur Einschätzung der Homogenität oder Heterogenität der Daten. [220]

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Formel 2 - Standardabweichung

3.6.2 Weitere Methoden

Zur Überprüfung der Zuverlässigkeit der eingesetzten Skalen wurde der Realibitätswert Cronbach's Alpha berechnet. Dieses Verfahren dient der Beurteilung der internen Konsistenz eines Fragebogenkonstrukt's, das aus mehreren Items besteht. Ein hoher Wert von über 0,7 wird üblicherweise als Indikator für eine hohe interne Konsistenz angesehen, wohingegen niedrigere Werte z. B. um 0,5 auf eine eher schwache Korrelation der Items hindeuten können. Ursachen für eine geringe interne Konsistenz können unter anderem unklare Formulierungen, zu wenige Items oder ein unzureichender inhaltlicher Zusammenhang zwischen den Items

sein. Cronbach's Alpha (α) trägt somit entscheidend dazu bei, die Reliabilität der Messinstrumente in dieser Arbeit sicherzustellen. [221]

$$\alpha = \frac{n}{n-1}(1 - \frac{\sum s^2(X_i)}{s^2(Y)})$$

Formel 3 - Cronbach Alpha

Zudem wurde ein weiterer statistischer Kennwert berechnet, um die Verteilung der Daten näher zu charakterisieren. Die Schiefe (Skewness) beschreibt dabei die Asymmetrie einer Verteilung. Positive Werte deuten auf Ausreißer auf der rechten Seite, negative Werte auf Ausreißer auf der linken Seite hin, während ein Wert von null eine symmetrische Verteilung signalisiert. Ist die Schiefe positiv (> 0), so ist die Verteilung rechtsschief. Ist die Schiefe negativ (< 0), so ist die Verteilung linksschief. [222]

$$\frac{\sum_{i=1}^{N}(X_i - \bar{X})^3}{(N-1)s^3}$$

Formel 4 - Schiefe

Kurtosis ist ein Maß dafür, wie stark eine Verteilung im Vergleich zur Normalverteilung zugespitzt oder abgeflacht ist. Der Referenzwert für die Normalverteilung liegt bei 3 und wird als mesokurtisch bezeichnet. Eine mesokurtische Verteilung entspricht damit in ihrer Form der Normalverteilung und dient als Maßstab, um Abweichungen nach oben oder unten einzuordnen. Liegt der Wert über 3, handelt es sich um eine leptokurtische Verteilung. Diese ist stark zugespitzt, die Werte konzentrieren sich stärker um den Mittelwert und gleichzeitig treten dicke Ränder auf, was bedeutet, dass extreme Ausreißer häufiger vorkommen. Liegt der Wert unter 3, spricht man von einer platykurtischen Verteilung, die flacher und breiter verläuft, eine gleichmäßigere Streuung zeigt und weniger anfällig für Ausreißer ist. Damit beschreibt Kurtosis nicht nur die Form des Zentrums einer Verteilung, sondern vor allem auch die Ausprägung der Enden, also der sogenannten Tails. Die Ergebnisse von Pearson verdeutlichen, dass reale Datensätze nur selten exakt mesokurtisch sind, sondern meist Tendenzen in Richtung leptokurtisch oder platykurtisch aufweisen. Somit ist Kurtosis ein wichtiges Instrument, um zu beurteilen, ob Daten eher ausreißeranfällig oder gleichmäßiger verteilt sind. [223]

$$n * \frac{\sum_i^n (Y_i - \bar{Y})^4}{\sum_i^n (Y_i - \bar{Y}^2)^2}$$

Formel 5 – Kurtosis

4 Ergebnisse

Zu Beginn dieses Kapitels wird die Forschungsfrage in Erinnerung gerufen: Wie beeinflusst die Verwendung eines humanoiden Roboters im Vergleich zu einem Web-Chatbot mit RAG die Akzeptanz und die Benutzerfreundlichkeit im Bildungsbereich?

Um diese Frage zu beantworten, nahmen insgesamt 20 Studierende des Studiengangs Data Science und AI am Experiment teil. Die Teilnehmenden wurden zufällig auf zwei Gruppen verteil. Eine Gruppe interagierte mit dem humanoiden Roboter Nao, die andere mit einem Web-Chatbot, beide Systeme basierend auf derselben RAG-Pipeline. Vor Beginn der Interaktion wurde ein Wissenstest als Pretest durchgeführt. Während des Experiments erfolgte die eigentliche Interaktion mit dem jeweiligen System. Im Anschluss daran beantworteten die Studierenden einen Akzeptanzfragebogen sowie den SUS-Fragebogen, bevor ein abschließender Wissenstest als Posttest erhoben wurde.

In diesem Kapitel erfolgt darauf aufbauend die detaillierte Datenanalyse. Die im Online-Fragebogen erhobenen Informationen werden aufbereitet und interpretiert, um die Forschungsfrage und Hypothesen fundiert beantworten zu können. Ziel ist es, belastbare und aussagekräftige Ergebnisse zu gewinnen. Dazu wurden die erhobenen Rohdaten von qualitativen in quantitative Merkmale überführt, sodass sie für die statistische Auswertung nutzbar sind (vgl. Tabelle 3). Für die Integration der Fragenpakete in die Analyse wird jeweils der Mittelwert gebildet und den entsprechenden Kategorien zugeordnet. Auf diese Weise lassen sich unter anderem die durchschnittliche Akzeptanz in Abhängigkeit von Faktoren wie der Erfahrung der Teilnehmenden untersuchen und mögliche Zusammenhänge zwischen den Variablen sichtbar machen. Die Analyse wird systematisch dokumentiert und transparent aufbereitet, um sowohl die Nachvollziehbarkeit als auch die Reproduzierbarkeit der Ergebnisse zu gewährleisten. Die Mittelwerte (vgl. Tabelle 14, Tabelle 15) von AF4 und AF5 dienen als Orientierungshilfe für die korrekte Anwendung des Allgemeinen Frage Konstrukts.

Mittelwert der Vertrautheit (AF4)	**Vertrautheit**
1,00-1,49	Überhaupt nicht vertraut
1,50-2,49	Wenig vertraut
2,50-3,49	Mäßig vertraut
3,50-4,49	Vertraut
4,50-5,00	Sehr vertraut

Tabelle 14 - Gruppierung der Vertrautheit (AF4)

Mittelwert der Nutzung (AF5)	Nutzung
1,00-1,49	Nie
1,50-2,49	Selten
2,50-3,49	Gelegentlich
3,50-4,49	Häufig
4,50-5,00	Sehr häufig / täglich

Tabelle 15 - Gruppierung der Nutzung (AF5)

4.1 Grundlegende Information zu den Proband*innen

An der Studie nahmen insgesamt 20 Proband*innen teil, die gleichmäßig auf die zwei verschiedenen Testsysteme NAO und Web-Chatbot verteilt wurden. Für jedes System wurden jeweils zehn Personen eingesetzt. Die Geschlechterverteilung war in beiden Gruppen ausgewogen. In der Web-Chatbot-Gruppe nahmen fünf männliche und fünf weibliche Personen teil, während die NAO-Gruppe aus sechs männlichen und vier weiblichen Teilnehmenden bestand. Diese ausgewogene Zusammensetzung trägt wesentlich zur Vergleichbarkeit der beiden Systemgruppen bei und bildet somit eine solide Grundlage für die anschließenden Analysen. (vgl. Tabelle 16)

System	Geschlecht	Häufigkeit
Web-Chatbot	**Männlich**	5
	Weiblich	5
	Divers	0
NAO	**Männlich**	6
	Weiblich	4
	Divers	0

Tabelle 16 - Häufigkeitstabelle des Geschlechts der Teilnehmer*innen

Die Altersverteilung der insgesamt zwanzig Befragten zeigt ein sehr ähnliches Profil in beiden Geschlechtergruppen. Von den elf männlichen Teilnehmenden ist circa ein Drittel (36,4 %) 20 Jahre alt, jeweils zwei Männer (18,2 %) sind 22 Jahre alt, einer (9,1 %) ist 23 Jahre alt, und weitere vier (36,4 %) fallen in die Sammelkategorie „25 Jahre und älter“. Jüngere Altersstufen unter 20 Jahren sowie der Einzelwert 24 Jahre sind bei den Männern nicht vertreten. Unter den neun weiblichen Befragten zeigt sich ein vergleichbares Bild. Drei Student*innen (33,3 %) sind 20 Jahre alt, zwei (22,2 %) gehören zur Kategorie 24 Jahre, jeweils eine (11,1 %) ist 19 bzw. 23 Jahre alt, und zwei (22,2 %) liegen in der Gruppe „25 Jahre und älter“. Damit

konzentriert sich der Großteil beider Gruppen auf das typische Studierendenalter zwischen 20 und 23 Jahren, während die Sammelkategorie 25+ auf eine kleinere, aber sichtbare Zahl älteren Studierenden hinweist. (vgl. Tabelle 17)

Geschlecht	Alter der Nutzer*innen	Abs. Häufigkeit	Rel. Häufigkeit (%)
Männlich	**19**	0	0,00
	20	4	36,36
	21	0	0,00
	22	2	18,18
	23	1	9,09
	24	0	0,00
	25+	4	36,36
	Total	**11**	**100**
Weiblich	**19**	1	11,11
	20	3	33,33
	21	0	0,00
	22	0	0,00
	23	1	11,11
	24	2	22,22
	25+	2	22,22
	Total	**9**	**100**

Tabelle 17 - Häufigkeitstabelle des Alters nach Geschlecht der Teilnehmer*innen

4.2 Allgemeine Fragen

Beide Geschlechter geben eine über der Skalenmitte liegende Vertrautheit (AF4) (vgl. Tabelle 18) an. Der Mittelwert der Vertrautheit liegt bei den Frauen bei 3,727, während die Männer einen Durchschnittswert von 3,778 erreichen. Damit positionieren sich die Befragten im Mittelwert von vertraut. Die Standardabweichungen fallen mit 0,647 (männlich) bzw. 0,667 (weiblich) sehr ähnlich aus. Die Spannweite unterscheidet sich jedoch deutlich zwischen den Geschlechtern. Bei den Männern reicht das Spektrum von „mässig vertraut“ bis „sehr vertraut“,

wobei ein Minimum unterhalb der neutralen Mitte nicht gewählt wurde. Bei den Frauen hingegen liegt der Bereich zwischen „wenig vertraut“ und „vertraut“.

Die fünfte allgemeine Frage (AF5) erfasst, wie häufig die Befragten digitale Lernmittel etwa Lernplattformen, Lernvideos oder KI-gestützte Tutorien einsetzen (vgl. Tabelle 19). Mit $\bar{x}$ von 4,364 geben männliche Befragte an, digitale Lernangebote im Schnitt „häufig“ zu nutzen. Weibliche Teilnehmende verorten sich ebenfalls mit einen Mittelwert von 3,778 an häufig. Die Standardabweichung ist bei Frauen mit 0,833 größer als bei Männern mit 0,674. Während einige Proband*innen digitale Lernmittel intensiv einsetzen, geben andere deutlich seltenere Nutzung an (bis hin zu wenig vertraut). Bei den Männern reicht das Spektrum nur von gelegentlich bis sehr häufig, niemand wählt die Kategorien „selten“ oder „nie“.

In beiden Untersuchungsgruppen (vgl. Abbildung 7,

Tabelle 20), dem humanoiden Roboter NAO und dem Web-Chatbot, zeigte sich nahezu

Systemtyp	**Häufigkeit**	**Min**	**$\bar{x}$**	**Max**	**σ**
NAO	10	3,00	4,20	5,00	0,632
Web-Chatbot	10	2,00	4,00	5,00	0,943
Total	**20**	**2,00**	**4,100**	**5,00**	**0,788**

derselbe Vertrautheitswert der Teilnehmenden im Umgang mit KI. Jeweils sieben von zehn Proband*innen des Roboters stuften sich selbst als „vertraut“ ein, während sich zwei weitere als „mäßig vertraut“ bezeichneten. Die Extremkategorien traten nur vereinzelt auf und verteilten sich gegengleich. Beim NAO fand sich eine Person, die „überhaupt nicht vertraut“ ist, wohingegen diese Stufe beim Chatbot nicht vorkam, umgekehrt erreichte nur beim Chatbot eine Versuchsperson die höchste Stufe „sehr vertraut“, die in der Gruppe des NAO fehlte. Die Kategorie „wenig vertraut“ wurde von keiner Versuchsperson gewählt. Der webbasierte Chatbot erreicht mit einem Mittelwert von 3,90 eine etwas höhere Nutzungshäufigkeit als der NAO-Roboter mit 3,60. Gleichzeitig fällt die Standardabweichung beim Web-Chatbot mit 0,568 geringer aus, was auf eine homogenere Nutzung im Vergleich zu NAO (σ = 0,699) hinweist. Insgesamt liegt der Gesamtdurchschnitt bei 3,75, sodass beide Systeme überwiegend im Bereich „häufig“ genutzt werden.

Die Auswertung (vgl. **Fehler! Verweisquelle konnte nicht gefunden werden.**, Abbildung 8) der Variable AF5 erfasst, wie häufig die Versuchspersonen allgemein KI-basierte Anwendungen nutzen. Beide Versuchsgruppen bringen ein hohes Maß an praktischer KI-Nutzung mit. In der NAO-Gruppe gab niemand an, KI „nie“ oder „selten“ zu verwenden, lediglich eine Person nutzte KI „gelegentlich“. Die Mehrheit setzt KI hingegen „häufig“ (6 von 10 Personen) oder „sehr häufig“ (3 von 10) ein. Für den Web-Chatbot zeigt sich ein nahezu deckungsgleiches Bild. Auch hier dominiert die Kategorie „häufig“ (5 von 10) gefolgt von „sehr häufig“ (3 von 10). Ein Unterschied besteht lediglich am unteren Rand der Skala, eine Person gab an, KI „nie“ zu nutzen, während die Stufe „selten“ von niemandem gewählt wurde. Der NAO-Roboter erreicht mit einem Mittelwert von 4,20 den höchsten Wert, was im Bereich „häufig“ liegt. Der webbasierte Chatbot erzielt mit einem Mittelwert von 4,00 einen ähnlich hohen, jedoch leicht niedrigeren Wert. Auffällig ist, dass die Streuung beim Web-Chatbot (σ = 0,943) größer ist als beim NAO (σ = 0,632).

Die Ergebnisse zeigen, dass die Befragten insgesamt eine hohe Vertrautheit mit KI und digitalen Lernmitteln aufweisen. Unterschiede zwischen Geschlechtern und Systemgruppen fallen gering aus, wobei beide Gruppen die Nutzung digitaler Lernangebote im Bereich „häufig“ einordnen. Die Streuungen sind system- und geschlechtsspezifisch leicht unterschiedlich, bewegen sich jedoch insgesamt in einem vergleichbaren Rahmen.

Geschlecht	**Häufigkeit**	**Min**	**x̄**	**Max**	**σ**
Männlich	11	3,00	3,727	5,00	0,647
Weiblich	9	2,00	3,778	4,00	0,667
Total	**20**	**2,00**	**3,750**	**5,00**	**0,639**

Tabelle 18 - Vertrautheit mit KI nach Geschlecht

Geschlecht	**Häufigkeit**	**Min**	**x̄**	**Max**	**σ**
Männlich	11	3,00	4,364	5,00	0,674
Weiblich	9	2,00	3,778	5,00	0,833
Total	**20**	**2,00**	**4,100**	**5,00**	**0,788**

Tabelle 19 - Nutzung von digitalen Lernmitteln nach Geschlecht

Systemtyp	**Häufigkeit**	**Min**	**x̄**	**Max**	**σ**
NAO	10	2,00	3,60	4,00	0,699
Web-Chatbot	10	3,00	3,90	5,00	0,568
Total	**20**	**3,00**	**3,750**	**2,00**	**0,639**

Tabelle 20 - Vertrautheit mit KI nach Untersuchungsgruppe

Systemtyp	**Häufigkeit**	**Min**	**x̄**	**Max**	**σ**
NAO	10	3,00	4,20	5,00	0,632
Web-Chatbot	10	2,00	4,00	5,00	0,943
Total	**20**	**2,00**	**4,100**	**5,00**	**0,788**

Tabelle 21 - Nutzung von digitalen Lernmitteln nach System

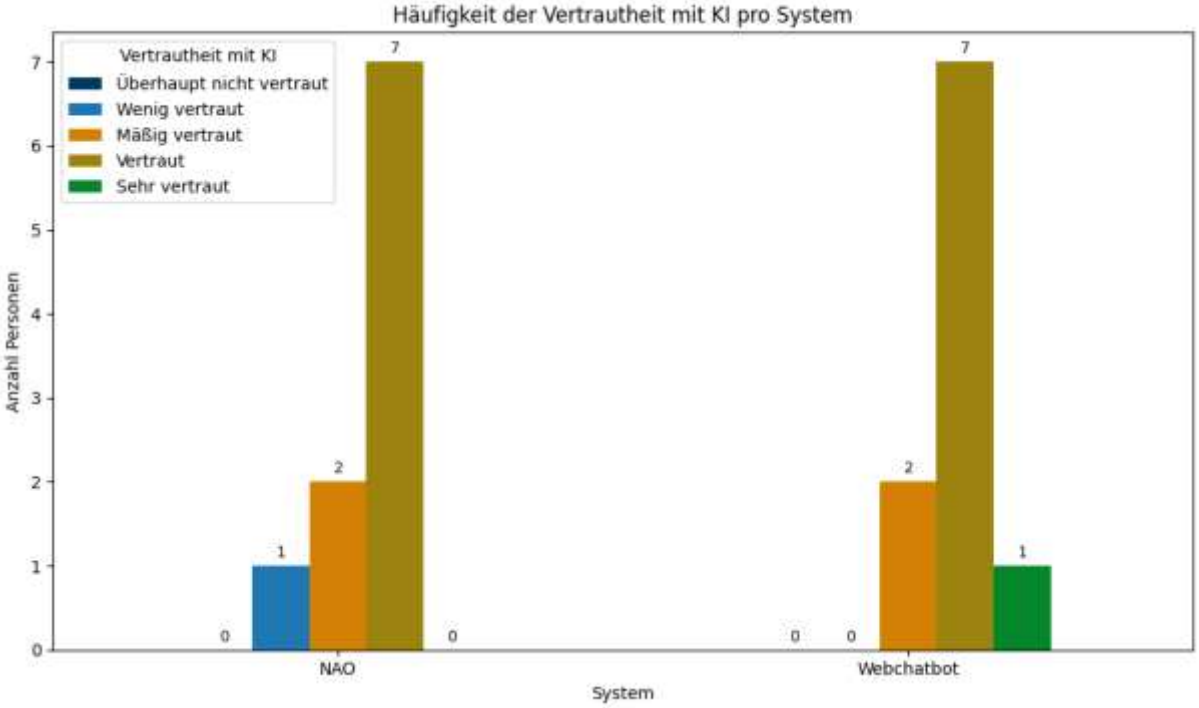

Abbildung 7 - Häufigkeit der Vertrautheit mit KI pro System

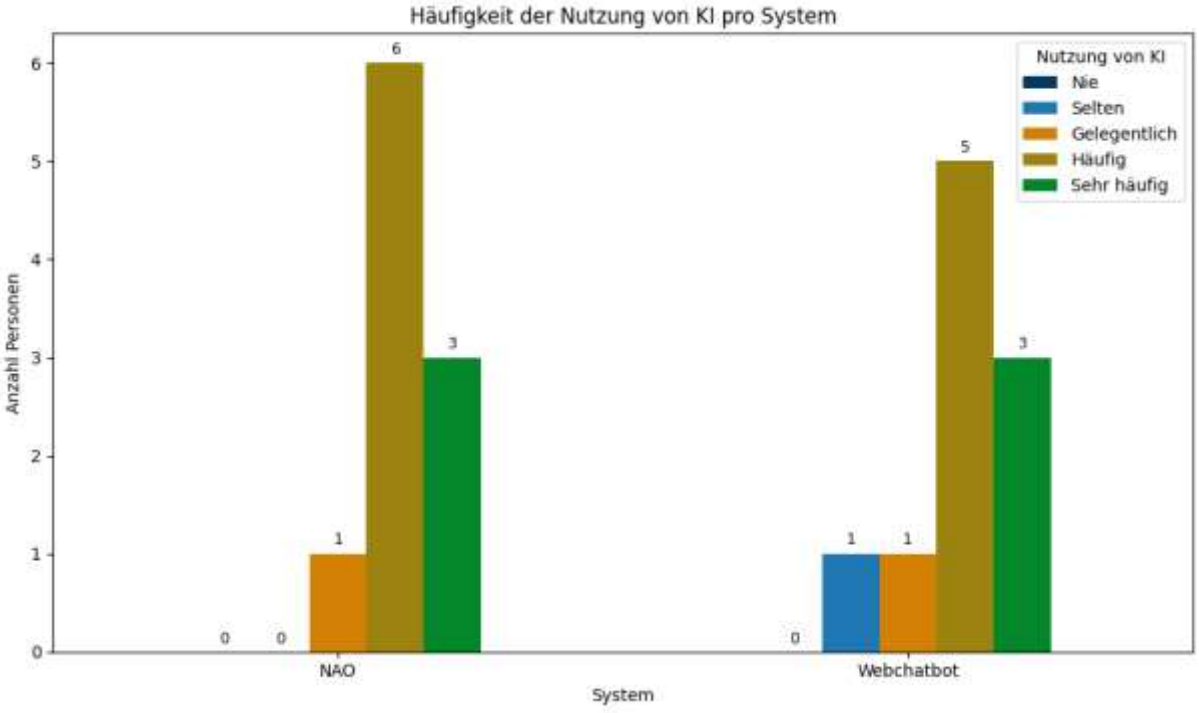

Abbildung 8 - Häufigkeit der Nutzung von KI pro System

4.3 Lernergebnisse

Die Auswertung der Lernergebnisse zeigt deutliche Unterschiede zwischen den beiden getesteten Systemen, dem humanoiden Roboter NAO und dem webbasierten Chatbot mit RAG. Im Pretest erreichen die NAO Proband*innen im Mittelwert 2,04 Punkte pro Frage, während der Web-Chatbot mit einem Mittelwert von 1,48 Punkten abschneidet. Die Standardabweichung liegt bei NAO bei 2,375 und beim webbasierten Chatbot bei 2,112. Betrachtet man die Summe dann erreicht NAO 102 Punkten und der Web-Chatbot 74 Punkten. In der Betrachtung der Summe wurden 176 Punkte erreicht.

Im Posttest erreichen die NAO Benutzer*innen einen Mittelwert von 1,76 Punkten, was einen klaren Abfall im Vergleich zum Pretest darstellt. Der Web-Chatbot lag im Posttest bei einem Mittelwert von 1,72 Punkten. Damit näherten sich die beiden Systeme deutlich an, und der Vorteil des Web-Chatbots aus dem Pretest wurde reduziert. Die Standardabweichungen lagen bei NAO auf 2,276 und beim Webchatbot auf 2,191. Die Gesamtsummen im Posttest zeigen, dass NAO mit 86 Punkten und der Web-Chatbot mit 88 Punkten nahezu gleich liegen. Die Gesamtsumme beider Systeme liegt bei 174 Punkten.

Zusammenfassend lässt sich festhalten, dass die Ergebnisse des Pretests höhere Mittelwerte für NAO zeigt. Im Posttest liegen beide Systeme mit nahezu identischen Mittelwerten und Gesamtsummen eng beieinander. Insgesamt gleichen sich die Unterschiede zwischen den beiden Systemen im Verlauf der Tests weitgehend an.

System	**Art des Tests**	**min**	**$\bar{x}$**	**max**	**σ**	**Summe**
Pretest	**NAO**	0	2,04	9	2,375	102
	Web-Chatbot	0	1,48	9	2,112	74
	Total	**0**	**1,76**	**9**	**44,567**	**176**
Posttest	**NAO**	0	1,76	9	2,276	86
	Web-Chatbot	0	1,72	9	2,191	88
	Total	**0**	**1,74**	**9**	**30,538**	**174**

Tabelle 22 - Deskriptive Statistik zu den Lernergebnissen

4.4 Akzeptanz (UTAUT)

Dieses Kapitel beleuchtet die Akzeptanz der beiden untersuchten Systeme in vier aufeinanderfolgenden Analyseschritten. Zunächst wird die allgemeine Akzeptanz jedes Systems betrachtet, um grundlegende Unterschiede in den Nutzerbewertungen sichtbar zu machen. Anschließend wird untersucht, ob sich diese Urteile in Abhängigkeit vom Geschlecht der Teilnehmenden verändern. Darauf folgt eine Analyse der Rolle der Vertrautheit mit Künstlicher Intelligenz, der Nutzungshäufigkeit digitaler Assistenten, die Datenverteilung und den Abschluss bildet eine Betrachtung der Überprüfung der Konsistenzreliabilität anhand vom Cronbach Alpha.

4.4.1 Allgemeine Akzeptanzanalyse nach System

Die deskriptiven Kennwerte der beiden UTAUT-Datensätze, einer für das humanoide Robotersystem NAO (vgl. Tabelle 23) sowie einer für den textbasierten Web-Chatbot (vgl. Tabelle 24), liefern ein detailliertes Bild der wahrgenommenen Akzeptanzfaktoren in den jeweiligen Gruppen. Sowohl beim Roboter als auch beim Chatbot ist die gesamte Antwortspanne von der minimalen bis zur maximalen Skalenstufe ausgewählt worden, was dafürspricht, dass die Items sensibel sind, um unterschiedliche Einstellungsstärken abzubilden und keine Decken- oder Bodeneffekte auftreten.

Beim NAO-System liegt der Mittelwert der Leistungserwartung bei 3,40 bei einer Standardabweichung von 1,32. Der Chatbot erreicht hier einen leicht höheren Wert von 3,55 bei einer Standardabweichung von 0,87. Beide Werte liegen oberhalb der neutralen Skalenmitte und weisen damit auf eine insgesamt positive Einschätzung hin. (vgl. Tabelle 23, Tabelle 24)

Die Aufwandserwartung zeigt in beiden Gruppen hohe Werte. Für NAO liegt der Mittelwert bei 3,90 mit einer Standardabweichung von 0,61. Beim Chatbot beträgt der Mittelwert 4,00, die Streuung liegt hier bei 1,38. (vgl. Tabelle 23, Tabelle 24)

Auch beim sozialen Einfluss liegen die Bewertungen nah beieinander. Der NAO erreicht einen Mittelwert von 3,33, der Chatbot 3,35. Unterschiede zeigen sich nur in den Minimalwerten, die beim NAO bei 1,00 und beim Chatbot bei 2,00 liegen. (vgl. Tabelle 23, Tabelle 24)

Für die erleichternden Bedingungen berichten beide Gruppen ein hohes Ausstattungs- und Infrastrukturlevel. Der NAO liegt hier bei einem Mittelwert von 3,97 mit einer Standardabweichung von 1,41. Der Chatbot erzielt 4,13 bei einer Standardabweichung von 0,67. (vgl. Tabelle 23, Tabelle 24)

Bezüglich der Verhaltensabsicht zeigen die Ergebnisse für den NAO einen Mittelwert von 3,23 (σ = 1,29) in der ersten Messung sowie 3,43 (σ = 1,36) in der zweiten Messung, bei einer Spannweite von 1,00 bis 5,00. Dies deutet darauf hin, dass die Nutzungsabsicht insgesamt im mittleren Bereich liegt, mit leichter Tendenz zu positiveren Bewertungen im Verlauf. (vgl. Tabelle 23, Tabelle 24)

Die Gesamtbewertung über alle UTAUT-Items liegt beim NAO bei 3,58 mit einer Standardabweichung von 1,16. Der Chatbot erreicht einen Wert von 3,69 bei einer Standardabweichung von 1,15. (vgl. Tabelle 23, Tabelle 24)

Insgesamt liegen die Bewertungen für beide Systeme in einem ähnlichen, leicht positiven Bereich. Der Chatbot erzielt in nahezu allen Dimensionen geringfügig höhere Mittelwerte als der Roboter, die Unterschiede bleiben jedoch klein. (vgl. Tabelle 23, Tabelle 24)

Fragenpaket	**min**	**x̄**	**max**	**σ**
LE	1,00	3,40	5,00	1,32
AE	2,00	3,95	5,00	0,61
SE	1,00	3,33	5,00	0,94
EB	1,00	3,97	5,00	1,41
VA	1,00	3,23	5,00	1,29
Allgemein	**1,00**	**3,58**	**5,00**	1,16

Tabelle 23 - Deskripitive Statistik - NAO (UTAUT)

Fragenpaket	**min**	**x̄**	**max**	**σ**
LE	1,00	3,55	5,00	0,87
AE	2,00	4,00	5,00	1,38
SE	2,00	3,35	5,00	1,05
EB	2,00	4,13	5,00	0,67
VA	1,00	3,43	5,00	1,36
Allgemein	**1,00**	**3,69**	**5,00**	1,15

Tabelle 24 - Deskripitive Statistik – Web-Chatbot (UTAUT)

4.4.2 Akzeptanzanalyse nach System und Geschlecht

Die Abbildung 9 zeigt die Leistungserwartung nach System und Geschlecht. Für den Roboter NAO bewerten die weiblichen Teilnehmenden die erwartete Leistungsfähigkeit mit einem Mittelwert von 3,69 höher als die männlichen Teilnehmenden, die 3,21 erreichen. Beim Web-Chatbot liegen die Werte näher beieinander. Männer erreichen hier 3,50, Frauen 3,60. Innerhalb der Geschlechtergruppen ergibt sich damit für Männer ein Zuwachs von fast drei Zehntelpunkten zugunsten des Chatbots. Frauen bewerten hingegen den Roboter mit 3,69 minimal besser als den Chatbot mit 3,60.

Die Abbildung 10 stellt die Aufwandserwartung dar. Beim Roboter NAO geben Männer mit 4,08 einen höheren Wert an als Frauen mit 3,75. Beim Web-Chatbot verringert sich der Unterschied, Männer bewerten den Aufwand mit 4,10, Frauen mit 3,90. Der Abstand reduziert sich damit auf 0,20 Punkte. Gleichzeitig verbessert sich die Bewertung der Frauen im Vergleich zum Roboter um 0,15 Punkte.

Die Abbildung 11 zeigt die Ergebnisse zum sozialen Einfluss. Beim Roboter NAO liegen Männer mit 3,38 leicht über den Frauen, die 3,25 erreichen. Frauen geben dem Web-Chatbot den höchsten Wert aller vier Gruppen mit 3,60. Männer bewerten den Chatbot dagegen mit 3,10. Der Chatbot-Wert der Frauen liegt damit 0,35 Punkte über ihrem NAO-Wert, während Männer den Chatbot um 0,28 Punkte niedriger einschätzen als den Roboter.

Die Abbildung 12 verdeutlicht die erleichternden Bedingungen. Beim Roboter NAO erreichen Frauen den höchsten Wert aller vier Gruppen mit 4,25. Männer liegen mit 3,78 deutlich darunter, die Differenz beträgt hier 0,47 Punkte. Beim Web-Chatbot dreht sich das Verhältnis um, Männer bewerten die erleichternden Bedingungen mit 4,20 am besten, Frauen liegen mit 4,07 knapp darunter daraus ergibt sich ein Unterschied von 0,13 Punkten kleiner.

Die Abbildung 13 zeigt die Verhaltensabsicht. Beim NAO liegen Männer mit 3,28 knapp über der Skalenmitte, Frauen erreichen mit 3,17 einen fast identischen Wert. Beim Web-Chatbot steigt die Verhaltensabsicht der Männer auf 3,80 an und liegt damit fast eine halbe Skalenstufe über dem NAO-Wert derselben Gruppe. Frauen bewerten den Chatbot mit 3,07 hingegen niedriger als den Roboter und unterschreiten damit als einzige Gruppe knapp den Bereich der positiven Zustimmung. Der Unterschied zwischen Männern und Frauen im Chatbot liegt damit bei 0,73 Punkten.

Die Abbildung 14 fasst die Gesamtwerte des UTAUT-Scores zusammen. Beim Roboter NAO erreichen die Männer einen mittleren Gesamtwert von 3,55, Frauen liegen mit 3,61 leicht darüber, die Differenz beträgt hier nur 0,06 Punkte. Beim Web-Chatbot zeigt sich ein umgekehrtes Muster, Männer erzielen mit 3,71 den höchsten Wert aller Gruppen, Frauen liegen mit 3,66 knapp darunter. Vergleicht man die Systeme innerhalb der Geschlechter, steigt der Gesamtwert der Männer vom Roboter zum Chatbot um 0,16 Punkte, bei den Frauen nur um 0,05 Punkte.

Die Ergebnisse zeigen, dass die Unterschiede zwischen den Geschlechtern insgesamt gering ausfallen. Beim Roboter liegen Männer und Frauen fast gleichauf, beim Chatbot erzielen Männer etwas höhere Werte, während Frauen leicht darunter liegen.

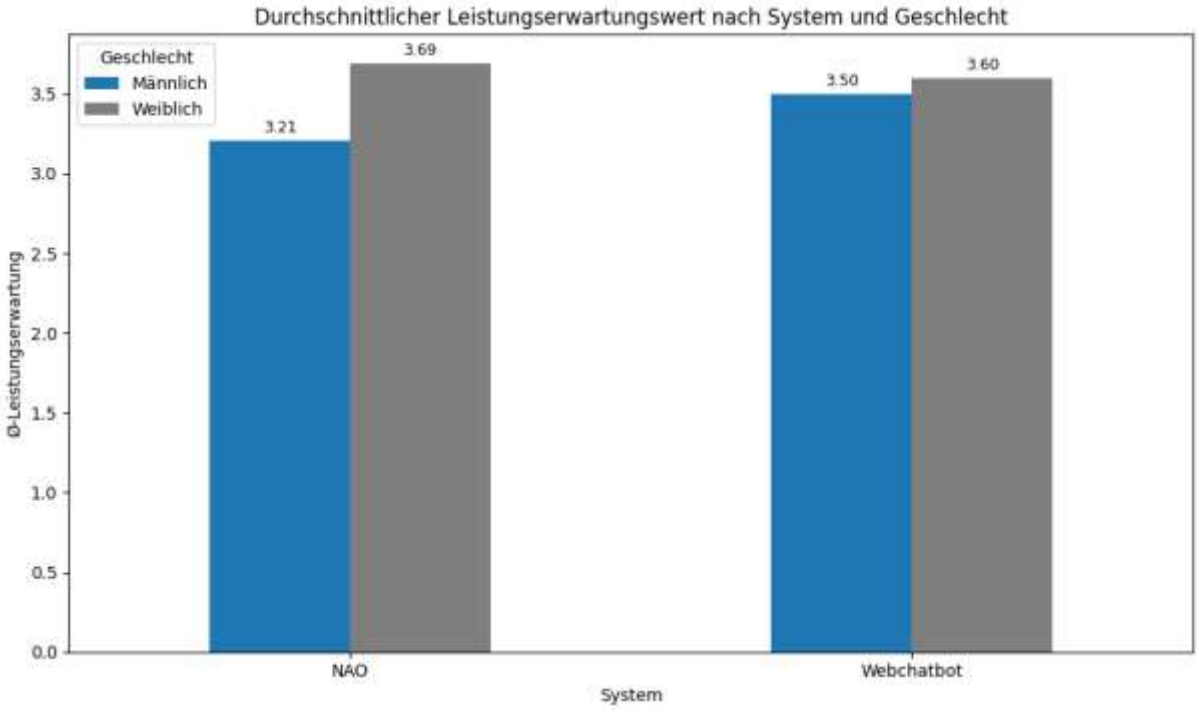

Abbildung 9 - Durchschnittlicher Leistungserwartungswert nach System und Geschlecht

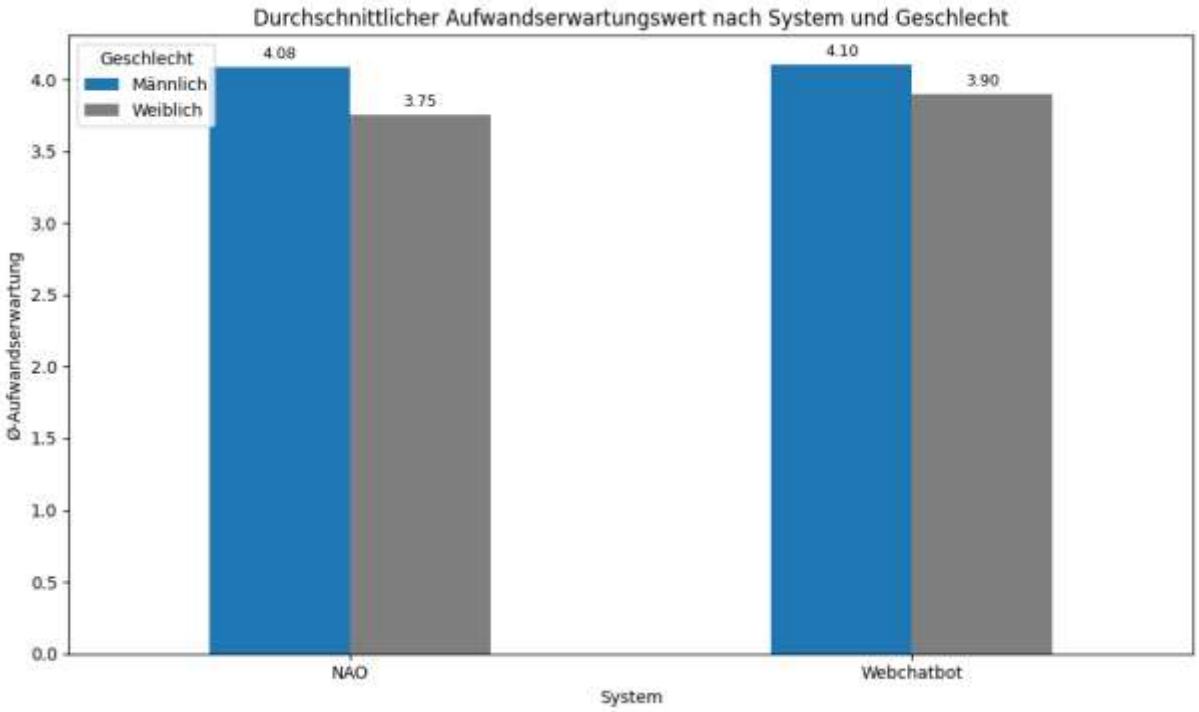

Abbildung 10 - Durchschnittlicher Aufwandserwartungswert nach System und Geschlecht

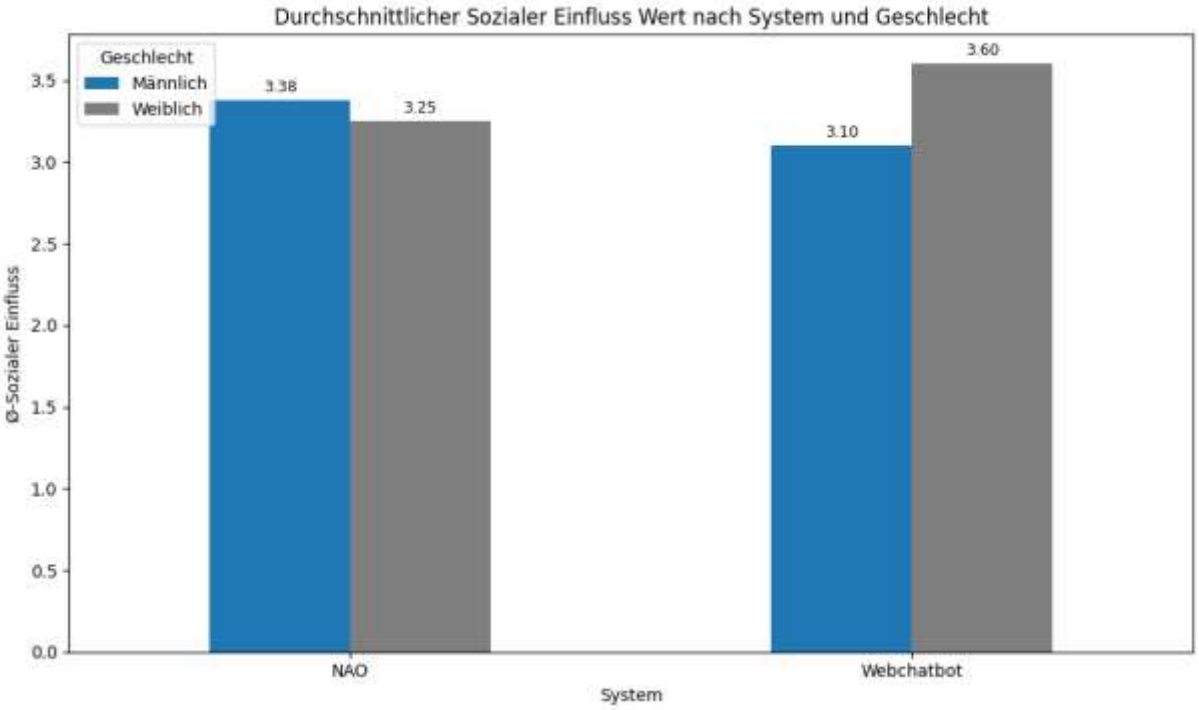

Abbildung 11 - Durchschnittlicher Sozialer Einfluss Wert nach System und Geschlecht

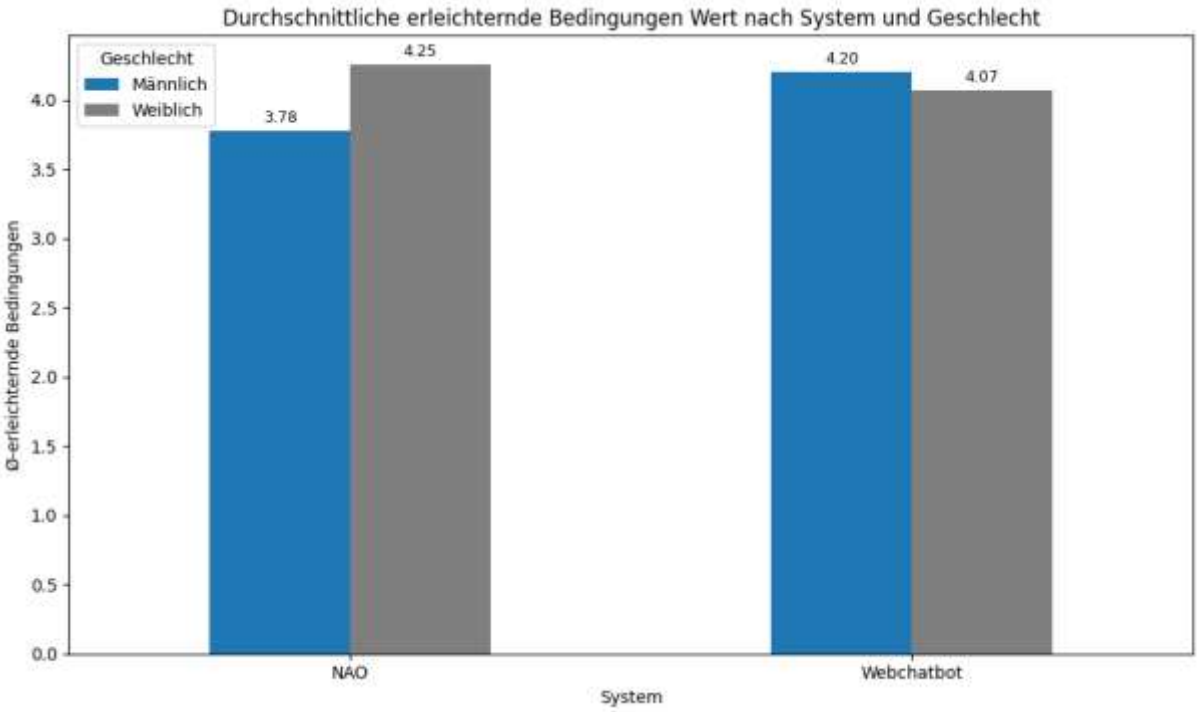

Abbildung 12 - Durchschnittliche erleichternde Bedingungen Wert nach System und Geschlecht

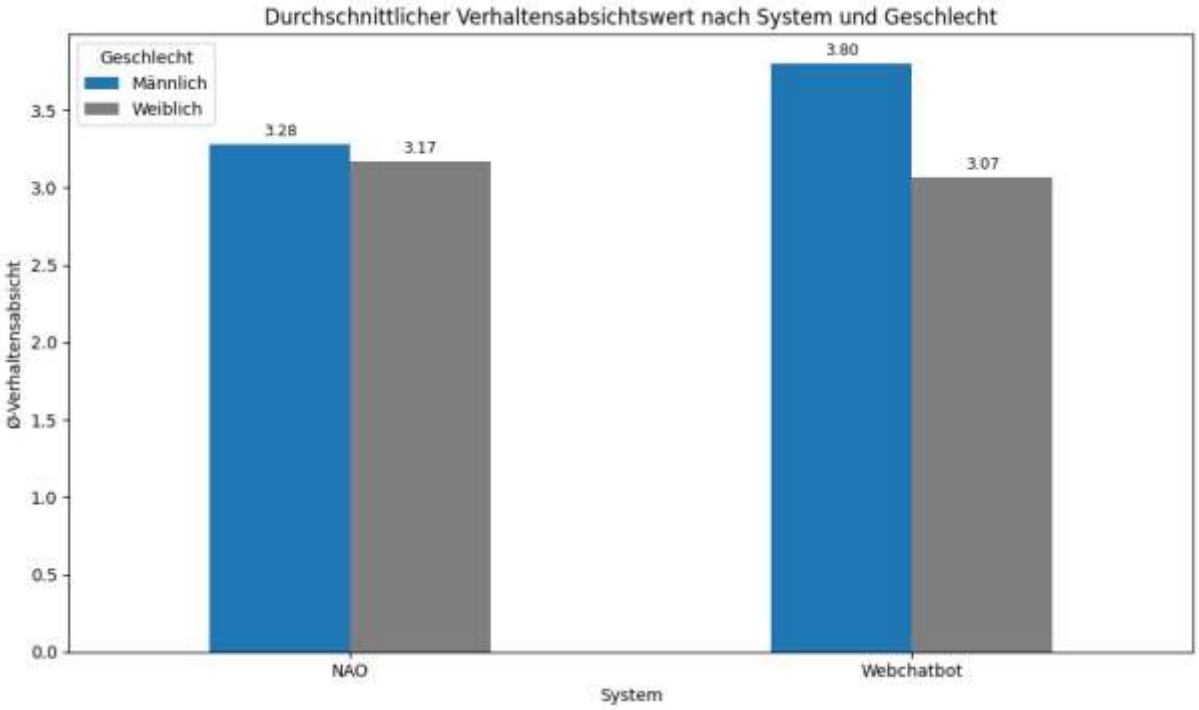

Abbildung 13 - Durchschnittlicher Verhaltensabsichtswert nach System und Geschlecht

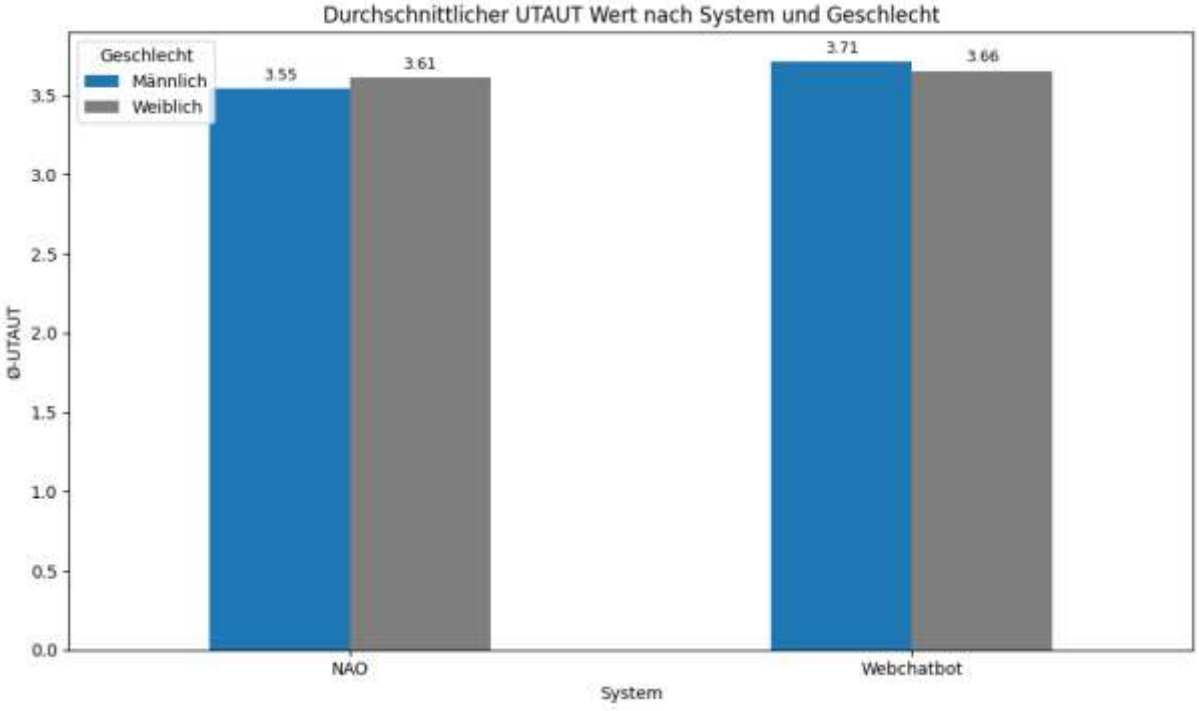

Abbildung 14 - Durchschnittlicher UTAUT Wert nach System und Geschlecht

4.4.3 Akzeptanzanalyse nach System und Vertrautheit von KI

Die Abbildung 15 zeigt die Leistungserwartung nach Vertrautheit mit KI. Der Roboter NAO wird in allen Gruppen als gleich leistungsfähig eingeschätzt. Teilnehmende mit geringer Erfahrung bewerten ihn mit einem Mittelwert von 3,50, Personen mit mittlerer Erfahrung mit 3,00 und die Gruppe der vertrauten Nutzenden erneut mit 3,50. Beim Web-Chatbot steigt die Leistungserwartung erst mit wachsender Expertise deutlich an, da mäßig vertraute Teilnehmende einen Wert von 2,75 angeben, während die Gruppe der Vertrauten mit 3,79 deutlich höher liegt. Sehr vertraute Proband*innen bewerten den Chatbot mit 3,50 und erreichen damit wieder das Niveau von NAO.

Die Abbildung 16 beschreibt, wie die Teilnehmenden den erforderlichen Aufwand beziehungsweise die wahrgenommene Aufwandserwartung, der beiden Systeme einschätzen. Für den humanoiden Roboter NAO ergibt sich ein leicht gekrümmtes Profil, Personen mit wenig KI-Expertise bewerten den die Aufwandserwartung mit einem Mittelwert von 3,75 auf der Likert-Skala, was bereits oberhalb des neutralen Bereichs liegt. Die Bewertung steigt bei der Gruppe mit mäßiger Vertrautheit auf 4,62, sinkt dann aber bei „vertrauten“ Nutzer*innen wieder auf 3,79. Mäßig erfahrene Personen geben dem Chatbot einen Wert von 4,25, damit bereits höher als NAO in derselben Erfahrungsstufe. Für „vertraute“ Nutzer*innen liegt der Wert mit 3,82 identisch zum Roboter, während die Gruppe „sehr vertraut“ mit 4,75 den höchsten Aufwandserwartungswert aller Messpunkte einschätzen.

Der Vergleich der UTAUT-Dimension „sozialer Einfluss“, zeigt dass alle Erfahrungsgruppen NAO mit Werten oberhalb von 3,2 auf der Likert-Skala bewerten, das Maximum liegt bei 3,62. Beim Web-Chatbot ist der soziale Einfluss deutlich heterogener. Während mäßig erfahrene und sehr erfahrene Nutzende mit 2,75 nur einen verhaltenen Gruppendruck wahrnehmen, steigt der Wert in der Erfahrungsstufe „vertraut“ auf 3,61 und erreicht damit fast das Niveau des Roboters. (vgl. Abbildung 17)

Die Abbildung 18 verdeutlicht die erleichternden Bedingungen. Hier liegen die Werte für beide Systeme deutlich oberhalb des Skalenmittelpunkts. Beim Roboter steigt die Bewertung mit zunehmender Vertrautheit von 3,67 auf 4,05. Der webbasierte Chatbot bewegt sich durchgängig in einem engen und hohen Bereich zwischen 4,00 und 4,19. Schon bei mäßiger Vertrautheit berichten die Teilnehmenden hier von spürbar komfortabler Unterstützung.

Die Messung der Verhaltensabsicht zeigt, wie stark Lernende beabsichtigen, das jeweilige System in Zukunft erneut zu nutzen oder aktiv in ihre Lernprozesse einzubinden. Beim humanoiden Roboter NAO steigt die Nutzungsabsicht zunächst von 3,00 bei wenig KI-erfahrenen Personen auf 3,67 in der Gruppe mit mäßiger Vertrautheit, fällt aber bei „vertrauten“ Anwender*innen auf 3,14 zurück. Der Web-Chatbot hingegen zeigt ein kontinuierliches Wachstum der Verhaltensabsicht mit zunehmender Expertise. Nutzer*innen mit mäßiger Erfahrung liegen mit 3,17 zunächst knapp über dem Wert des Roboters, doch bereits in der Stufe „vertraut“ steigt die Absicht auf 3,43 und erreicht bei „sehr vertrauten“ Personen den Spitzenwert von 4,00. (vgl. Abbildung 19)

Die aggregierten UTAUT-Scores (vgl. Abbildung 20) liefern ein zusammenfassendes Bild der Akzeptanz beider Systeme über alle untersuchten Konstrukte hinweg. Beim humanoiden Roboter NAO liegen die Mittelwerte in einem engen Feld zwischen 3,44 und 3,75 und überschreiten damit in allen Erfahrungsstufen den neutralen Mittelpunkt der Skala. Der

humanoide Roboter wird bereits von Lernenden mit nur geringer KI-Vertrautheit als akzeptabel wahrgenommen und erzielt bei mittelmäßig Vertrauten sogar den höchsten Wert. Der Web-Chatbot zeigt ein anderes Profil. Proband*innen mit mäßiger Erfahrung bewerten ihn mit 3,36 zunächst etwas zurückhaltender als NAO. Mit zunehmender Vertrautheit steigt der Gesamt-UTAUT jedoch bis auf 3,78 und erreicht damit den Spitzenwert der gesamten Stichprobe.

Die Ergebnisse zeigen, dass NAO in allen Gruppen konstant akzeptiert wird, während der Chatbot stärker von der Vertrautheit mit KI profitiert. Besonders vertraute und sehr vertraute Teilnehmende bewerten den Chatbot höher, während der Roboter bei geringer und mittlerer Erfahrung stabil bleibt.

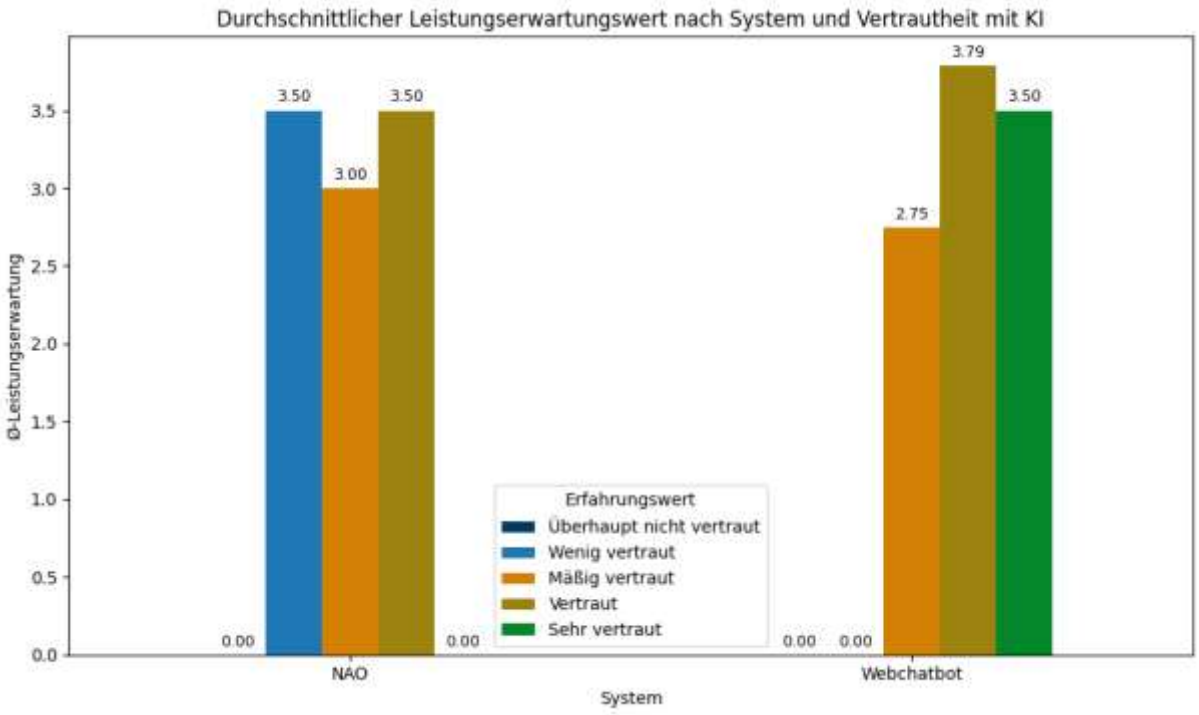

Abbildung 15 - Durchschnittlicher Leistungserwartungswert nach System und Vertrautheit mit KI

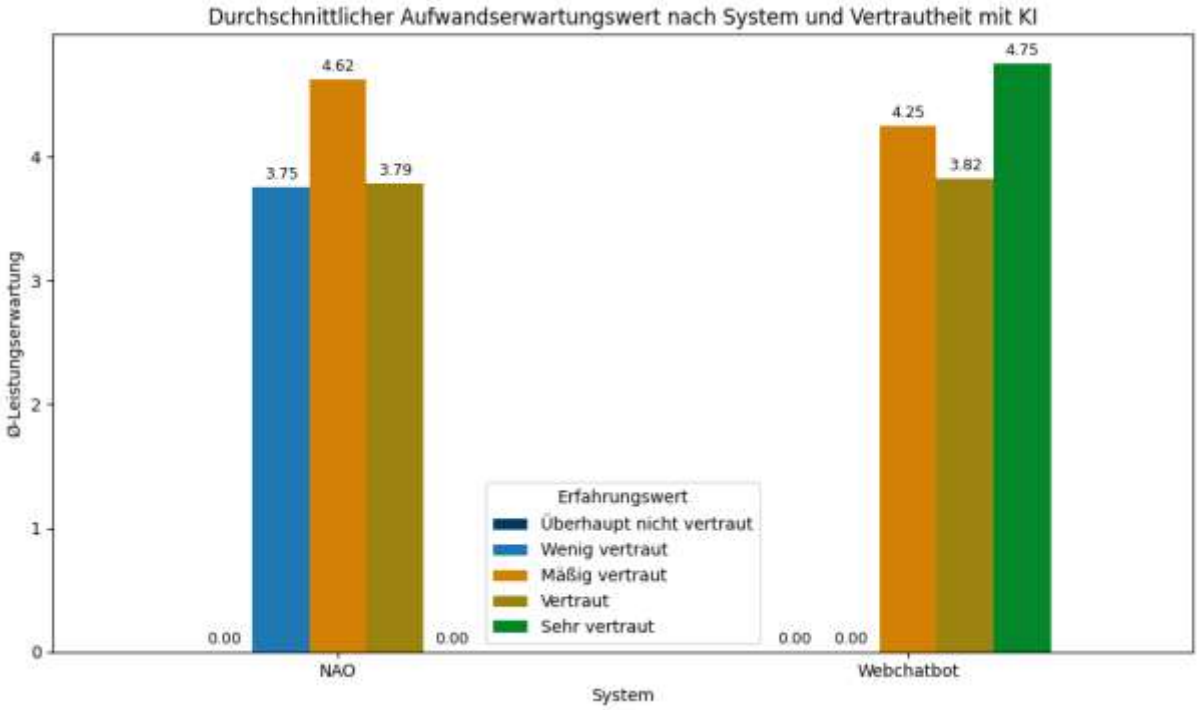

Abbildung 16 - Durchschnittlicher Aufwandserwartungswert nach System und Vertrautheit mit KI

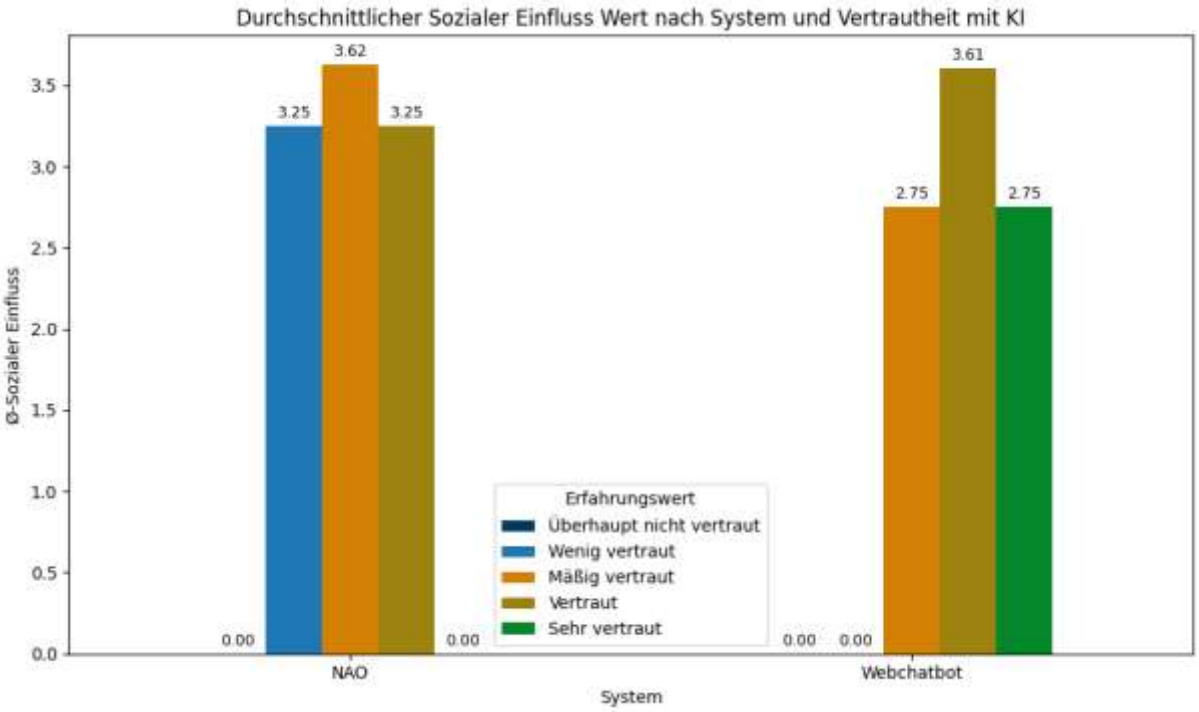

Abbildung 17 - Durchschnittlicher Sozialer Einfluss Wert nach System und Vertrautheit mit KI

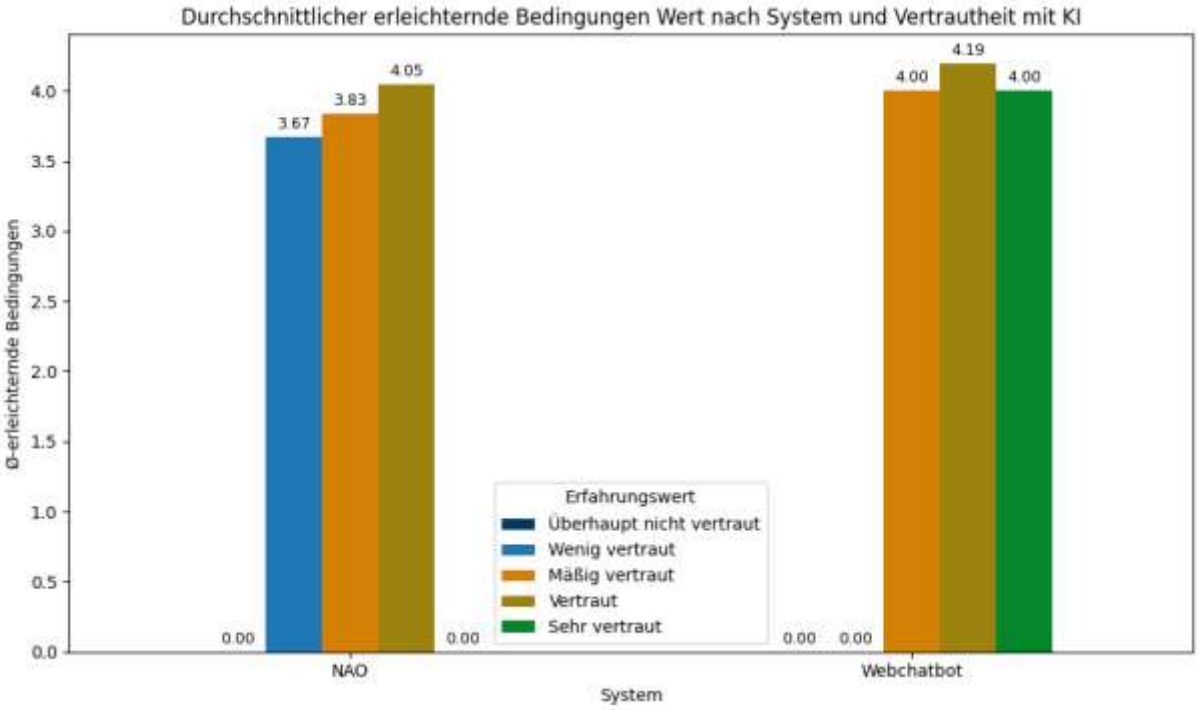

Abbildung 18 - Durchschnittlicher erleichternde Bedingungen Wert nach System und Vertrautheit mit KI

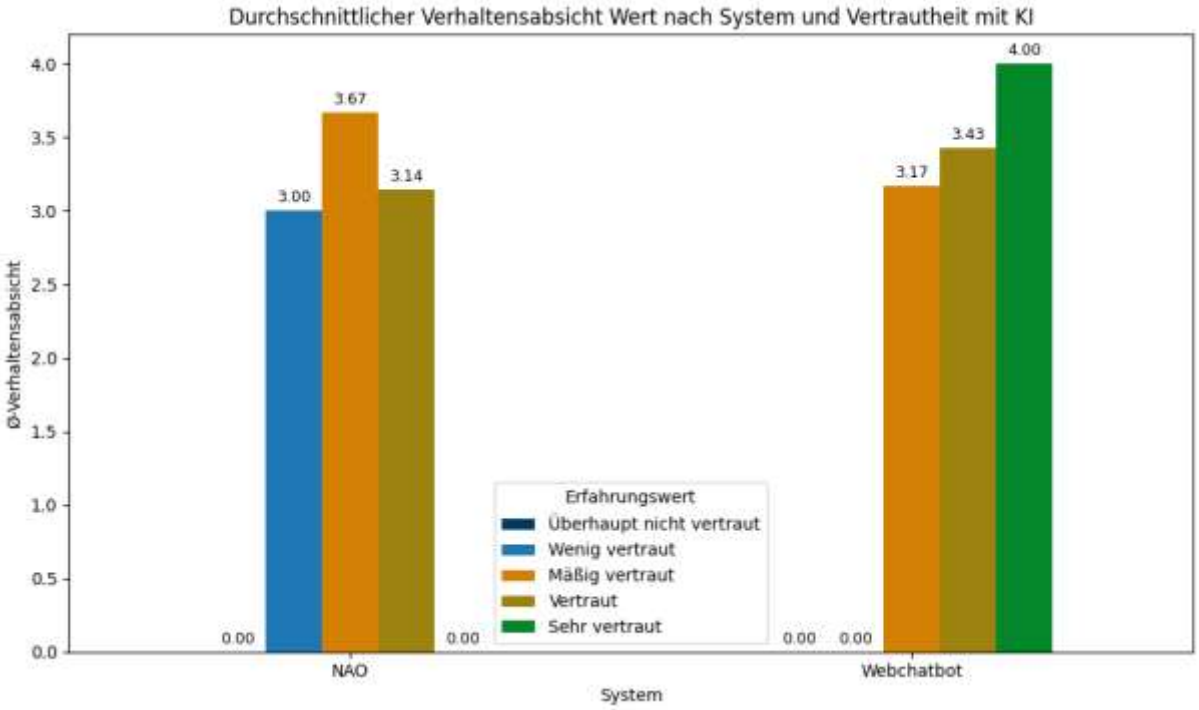

Abbildung 19 - Durchschnittlicher Verhaltensabsicht Wert nach System und Vertrautheit mit KI

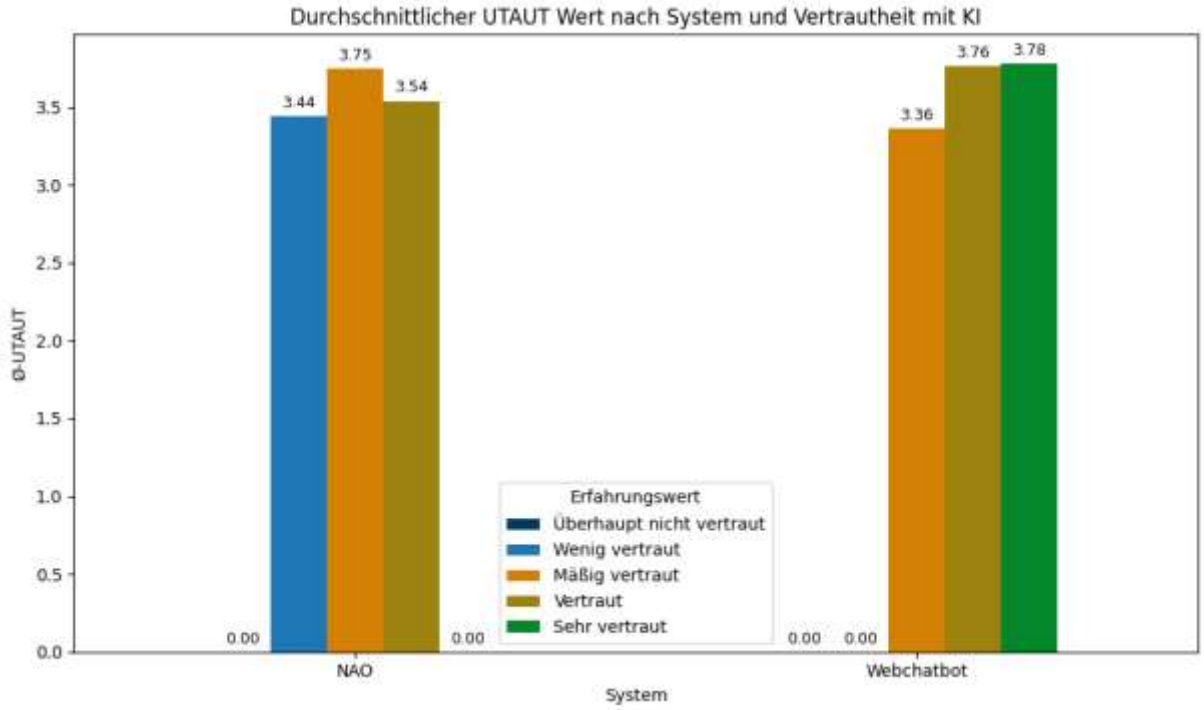

Abbildung 20 - Durchschnittlicher UTAUT Wert nach System und Vertrautheit mit KI

4.4.4 Akzeptanzanalyse nach System und Nutzungshäufigkeit

Die Abbildung 21 beschreibt die Leistungserwartung mit der Nutzungshäufigkeit von KI-Tools und zeigt zwei deutlich unterschiedliche Kurven. Beim humanoiden Roboter NAO erreicht die wahrgenommene Leistungsfähigkeit in der Gruppe der Teilnehmenden, die KI nur gelegentlich einsetzen, den Höchstwert von 4,00 Punkten. Mit zunehmender Routine sinkt dieser Wert schrittweise. Häufige Nutzer*innen bewerten NAO mit 3,50, sehr häufige Anwender*innen mit 3,00. Für den Web-Chatbot zeichnet sich das spiegelverkehrte Bild ab. Personen, die KI selten oder nie verwenden, bewerteten ihn mit 2,50 Punkten den niedrigsten Nutzen im gesamten UTAUT-Datensatz. Steigt die Nutzungshäufigkeit auf gelegentlich oder häufig, steigt die Leistungserwartung auf 3,75 beziehungsweise 3,95. Bei intensiver KI-Praxis bleibt der Wert mit 3,17 zwar leicht hinter der Häufig-Gruppe zurück, liegt aber weiterhin über dem Vergleichswert des Roboters. Zusammenfassend kann man behaupten, dass NAO bei der Leistungserwartung vor allem von Gelegenheitsnutzenden hoch bewertet wird und sinkt mit zunehmender Erfahrung, während der Web-Chatbot mit wachsender Nutzung deutlich gewinnt und bei intensiver Praxis über NAO liegt.

Die Aufschlüsselung der Aufwandserwartung nach Nutzungshäufigkeit von KI offenbart ein differenziertes Bild. Beim humanoiden Roboter NAO empfinden Lernende, die KI nur gelegentlich einsetzen, den Umgang als besonders mühelos, ihr Wert von 4,75 liegt deutlich über allen anderen Gruppen. Mit steigender Nutzung sinkt die wahrgenommene Leichtigkeit zunächst auf 3,92 bei häufigen und weiter auf 3,75 bei sehr häufigen Anwender*innen. Der Web-Chatbot zeigt ein gegenteiliges Muster. Personen, die KI selten nutzen, bewerten die Aufwandserwartung mit 3,75. In der Stufe „gelegentlich“ steigt der Wert auf 4,25, bevor er bei häufigen Nutzer*innen auf 3,70 zurückfällt. Sehr intensive Anwender*innen erreichen mit 4,50 den Spitzenwert. Zusammengefasst spricht das Muster dafür, dass NAO vor allem bei sporadischer Nutzung als besonders mühelos erlebt wird, die Aufwandserwartung mit

zunehmender Nutzung aber nachlässt, während der Web-Chatbot mit wachsender Routine leichter handhabbar wird und bei intensiver Nutzung die geringste Aufwandserwartung aufweist. (vgl. Abbildung 22)

Die Dimension „sozialer Einfluss“ macht deutlich, in welchem Maß das unmittelbare Umfeld die Nutzung des jeweiligen Systems befürwortet. Beim humanoiden Roboter NAO erreicht dieses Konstrukt seinen Höhepunkt in der Kohorte, die KI nur gelegentlich einsetzt. Mit einem Mittelwert von 4,00 Punkten signalisiert diese Gruppe, dass der Roboter in ihrem sozialen Umfeld als besonders wertvoll gilt. Wenn die Lernenden NAO häufiger oder sehr häufig verwenden, sinkt der soziale Einfluss auf 3,33 beziehungsweise 3,08. Der Web-Chatbot weist ein leicht versetztes Muster auf. Nutzer*innen, die KI selten verwenden, spüren mit 3,50 bereits einen moderaten, positiven Gruppendruck. In der Gruppe der „Gelegentlichen“ steigt der Wert ebenfalls auf 4,00 und erreicht damit denselben Spitzenwert wie NAO. Häufige Anwender*innen erleben einen Rückgang auf 3,20. Der soziale Einfluss steigt bei sehr häufiger Nutzung wieder auf 3,33. Die Befunde deuten darauf hin, dass der soziale Einfluss für beide Systeme vor allem bei gelegentlicher Nutzung am stärksten ist und mit zunehmender Nutzung abnimmt, beim Web-Chatbot jedoch mit einem leichten Wiederanstieg bei sehr intensiver Nutzung, was auf Gewöhnung bzw. eine bessere Einbindung ins Umfeld hindeutet. (vgl. Abbildung 23)

Die Abbildung 24 wahrgenommenen „erleichternde Bedingungen“ spiegeln wieder, ob Infrastruktur, technischer Support und organisatorische Rahmenbedingungen als ausreichend für den Systemeinsatz empfunden werden. Beim humanoiden Roboter NAO erreicht dieser Indikator mit einem Wert von 4,67 seinen Höhepunkt in der Gruppe, die KI gelegentlich nutzt. Mit zunehmender Nutzungshäufigkeit sinkt die Bewertung jedoch kontinuierlich auf 4,11 bei häufigen und 3,44 bei sehr häufigen Anwender*innen. Beim Web-Chatbot zeigt sich ein leicht anderes Bild. Personen, die KI überhaupt nicht oder nur gelegentlich einsetzen, beurteilen die erleichternden Bedingungen mit jeweils 4,67 außerordentlich positiv. Bei häufiger Nutzung sinkt der Wert auf 3,80 ab. Sehr intensive Nutzer*innen melden mit 4,33 jedoch erneut eine höhere Zufriedenheit.

Die Auswertung der Verhaltensabsicht (vgl. Abbildung 25), also der Bereitschaft, das System künftig aktiv einzusetzen, zeichnet erneut ein divergierendes Bild zwischen den beiden Systemen. Für NAO erreicht die Nutzungsabsicht in der Gruppe der Lernenden, die KI nur gelegentlich verwenden, mit 4,67 den höchsten Wert der gesamten Stichprobe. Sobald der Roboter jedoch regelmäßig genutzt wird, sinkt die Absicht deutlich auf 3,06 und stabilisiert sich bei sehr häufiger Nutzung bei 3,11. Beim Web-Chatbot zeigt sich ein kontinuierlicher Aufbau der Verhaltensabsicht über alle Nutzungsstufen hinweg. Lernende, die KI selten einsetzen, bewerten ihre künftige Chatbot-Nutzung zwar mit 2,00 Durchschnittspunkten, doch bei gelegentlicher Anwendung klettert der Wert auf 3,67, bleibt bei häufiger Nutzung mit 3,53 nahezu konstant und erreicht bei sehr intensiver Nutzung erneut 3,67.

Die aggregierten UTAUT-Werte (vgl. Abbildung 26) zeigen, wie die Akzeptanz beider Technologien von der individuellen Nutzungsfrequenz abhängt. Beim humanoiden Roboter NAO erreicht er die Durchschnittspunkte mit 4,39 ihr Maximum genau dann, wenn Lernende KI nur gelegentlich einsetzen. Sobald der Roboter hingegen häufiger oder sehr häufig zum Einsatz kommt, sinkt der UTAUT-Score auf 3,58 beziehungsweise 3,28. Anwender*innen die KI selten verwenden, vergeben ihm im Durchschnitt 3,28 Punkte. Bereits bei gelegentlicher Nutzung steigt der Wert jedoch deutlich auf 4,06 und übertrifft damit NAO. Bei häufiger

Anwendung fällt der Score leicht auf 3,63, klettert aber bei sehr intensiver Nutzung wieder auf 3,78. Die Ergebnisse zeigen, dass NAO vor allem bei gelegentlicher KI-Nutzung am besten bewertet wird, während seine Werte bei häufigerem Anwender*innen sinken. Der Web-Chatbot hingegen profitiert von zunehmender Nutzungserfahrung und erzielt bei intensiver Anwendung höhere Werte, besonders bei Verhaltensabsicht und Gesamtakzeptanz.

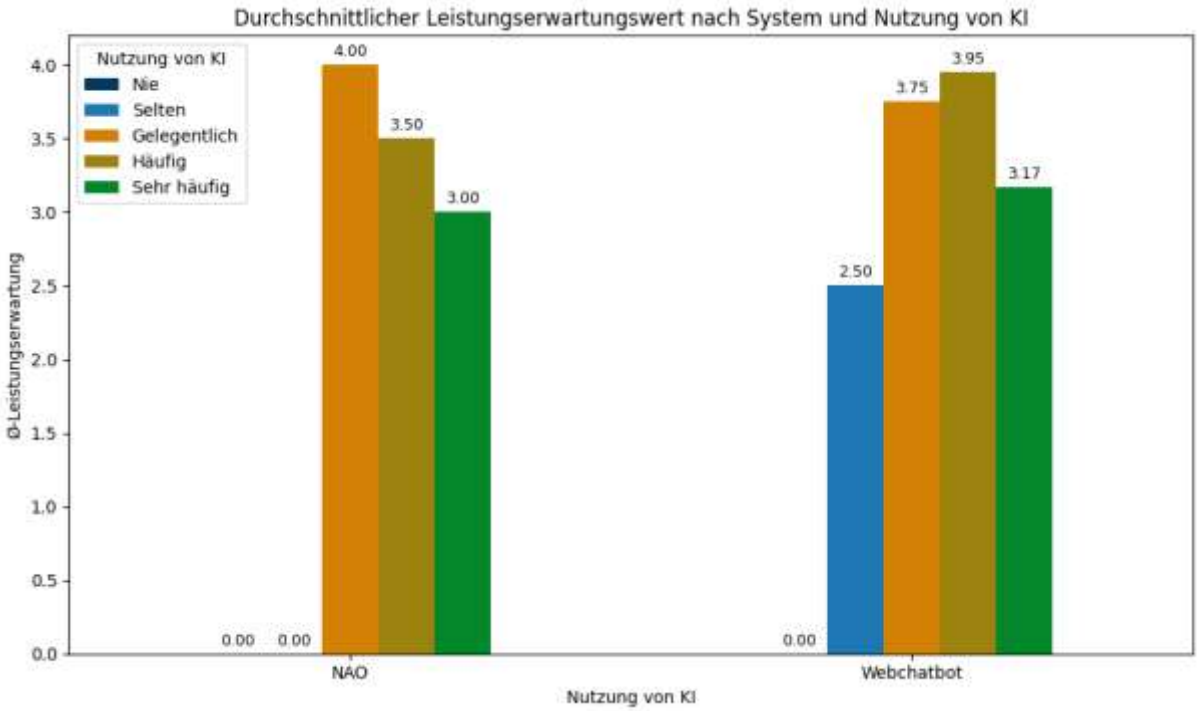

Abbildung 21 - Durchschnittlicher Leistungserwartungswert nach System und Nutzung von KI

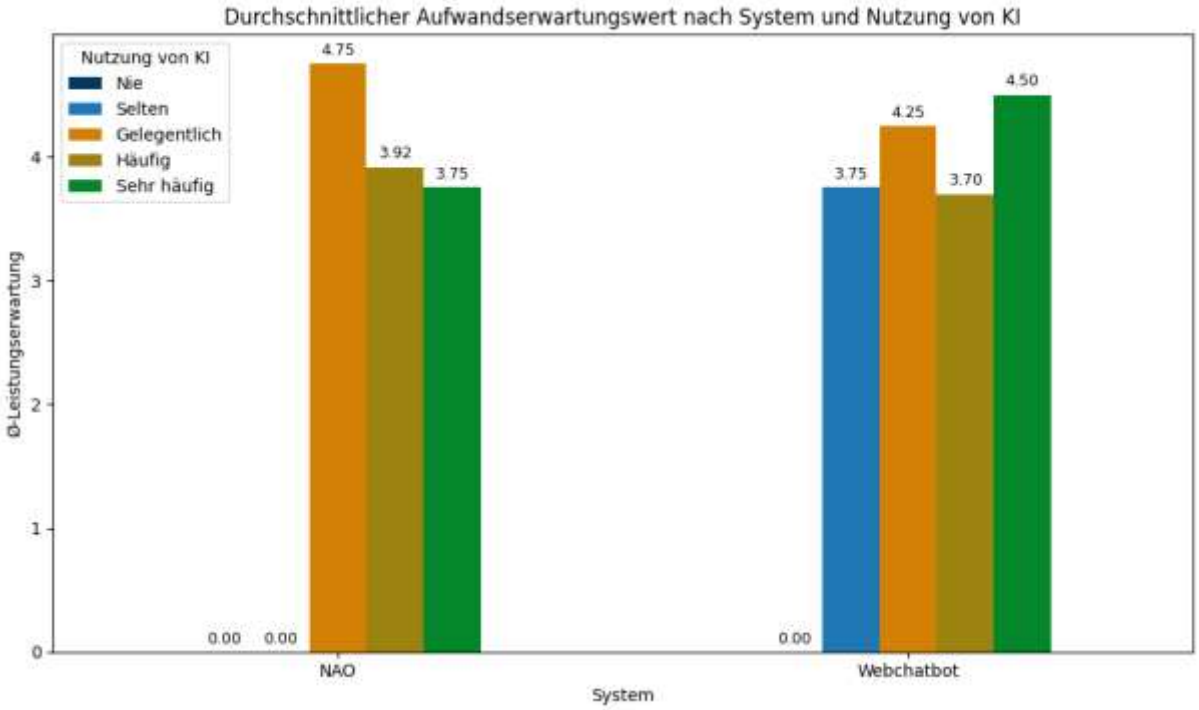

Abbildung 22 - Durchschnittlicher Aufwandserwartungswert nach System und Nutzung von KI

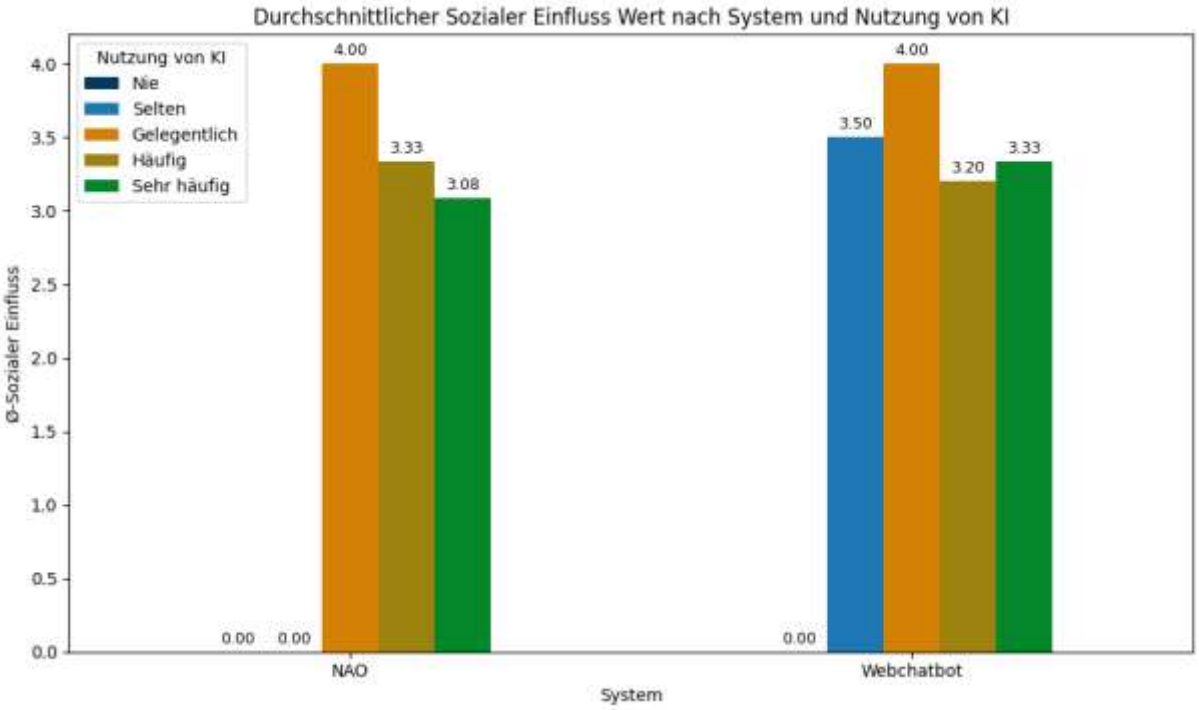

Abbildung 23 - Durchschnittlicher Sozialer Einfluss Wert nach System und Nutzung von KI

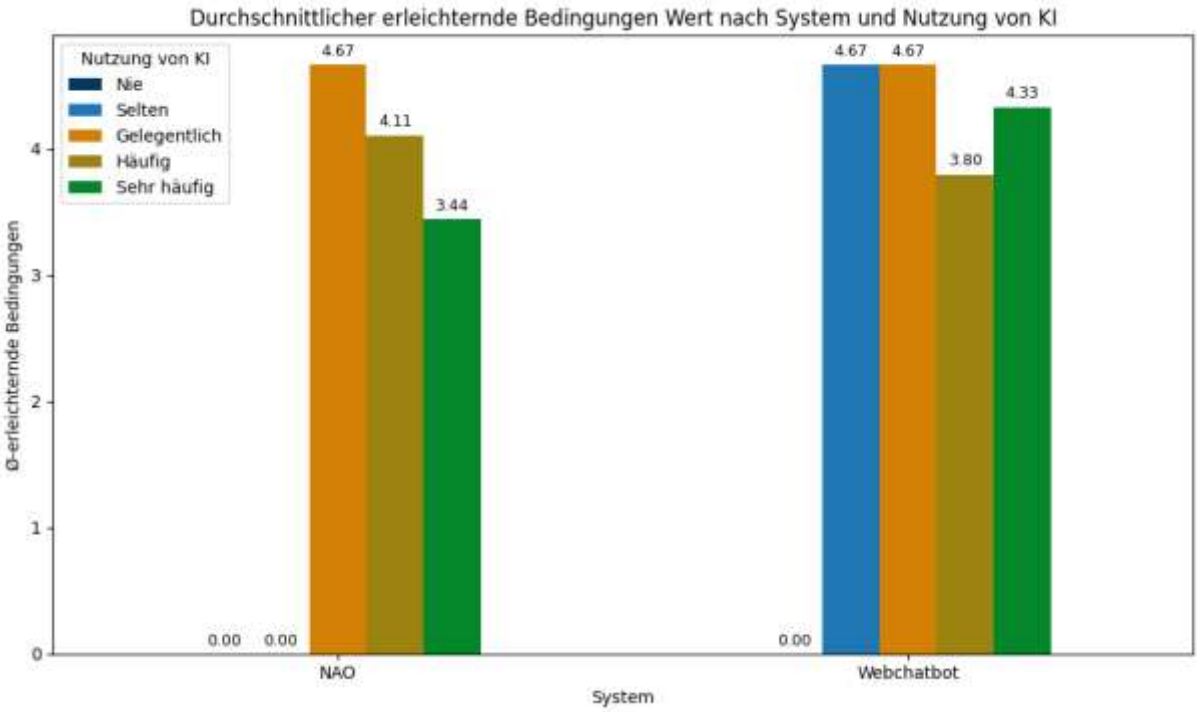

Abbildung 24 - Durchschnittlicher erleichternde Bedingungen Wert nach System und Nutzung von KI

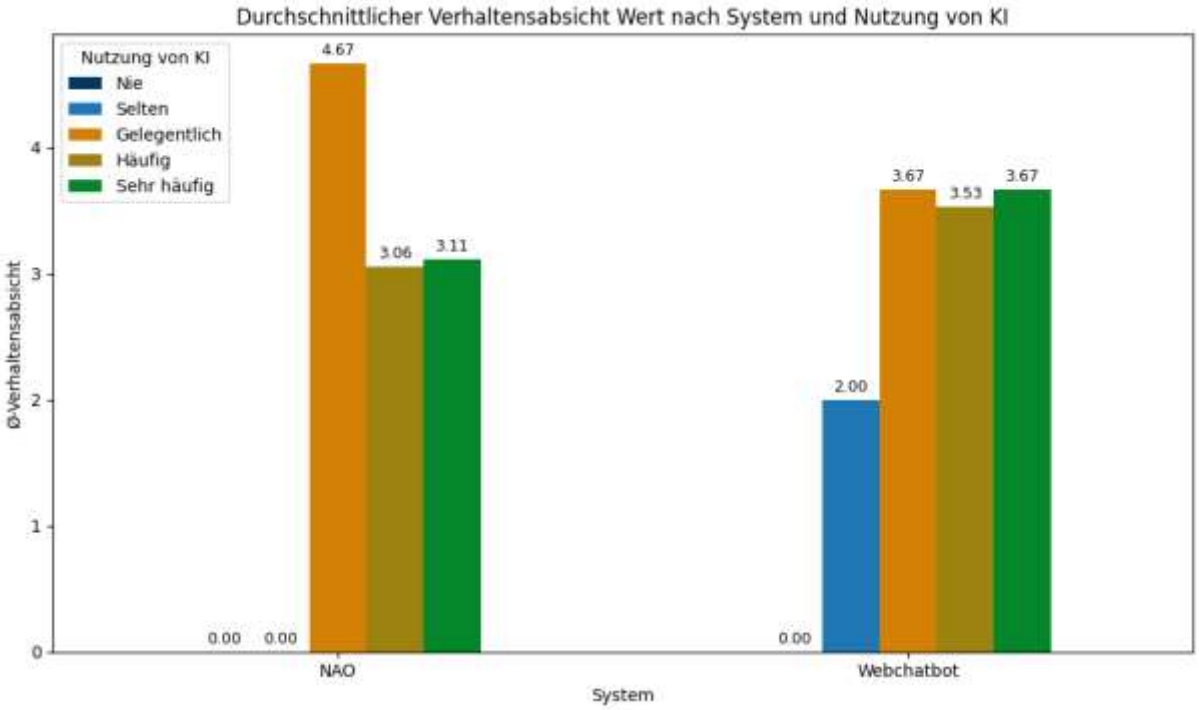

Abbildung 25 - Durchschnittlicher Verhaltensabsicht Wert nach System und Nutzung von KI

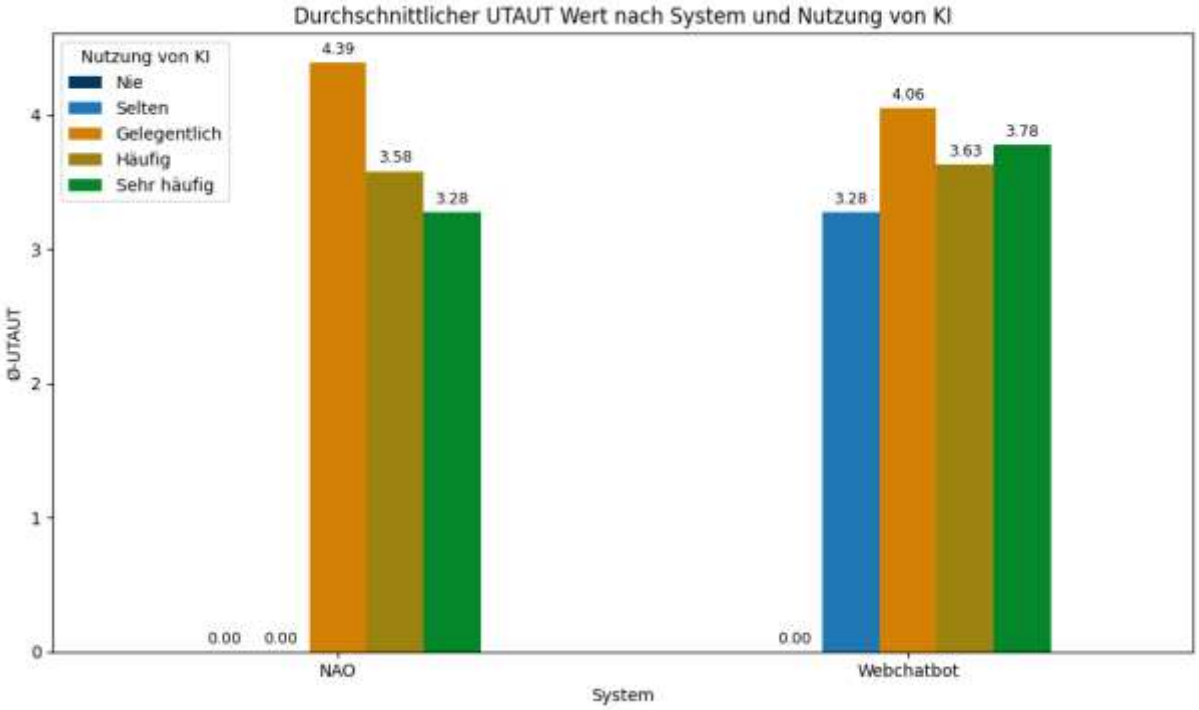

Abbildung 26 - Durchschnittlicher UTAUT Wert nach System und Nutzung von KI

4.4.5 Datenverteilung

Zur Analyse der Verteilungseigenschaften der erhobenen Daten wurden die Kennwerte Schiefe und Kurtosis berechnet. (vgl. Tabelle 25) Die Werte geben Hinweise darauf, wie die Antworten der Teilnehmenden um den Mittelwert verteilt sind und ob Abweichungen von einer Normalverteilung vorliegen.

Für die Skala Leistungserwartung zeigen sowohl der NAO als auch der webbasierte Chatbot negative Schiefewerte von -0,647 beziehungsweise -0,752. Dies deutet darauf hin, dass die Antworten tendenziell zu höheren Ausprägungen hin verschoben sind. Die Kurtosiswerte betragen -0,481 für NAO und 0,337 für den Web-Chatbot, was annähernd auf normalverteilte Antworten hinweist.

Bei der Aufwandserwartung weist NAO eine leicht negative Schiefe von -0,248 auf, während der Web-Chatbot mit -0,795 stärker zu hohen Werten tendiert. Die Kurtosis liegt bei -0,477 für NAO und -0,905 für den webbasierten Chatbot, was auf leicht abgeflachte Verteilungen hinweist.

Die Skala soziale Einflüsse zeigt beim NAO eine nahezu neutrale Verteilung mit einer Schiefe von -0,184 und einer Kurtosis von -0,375, während beim webbasierten Chatbot eine leicht positive Schiefe von 0,128 und eine stärker abgeflachte Verteilung von -1,091 vorliegt.

Bei den erleichternden Bedingungen ist die Schiefe für NAO mit -1,252 stark negativ, was auf eine deutliche Tendenz zu hohen Bewertungen hinweist. Die Kurtosis beträgt 0,938 und zeigt eine spitze Verteilung. Der Web-Chatbot zeigt hier eine moderate negative Schiefe von -0,662 und eine Kurtosis von -0,0262.

Die Verhaltensabsicht liegt bei beiden Systemen im negativen Bereich, NAO bei -0,493 und bei dem Web-Chatbot bei -0,809, was ebenfalls eine Tendenz zu höheren Bewertungen signalisiert. Die Kurtosiswerte betragen -0,349 für NAO und -0,0318 für den Webchatbot, was auf normalverteilte Antworten hindeutet.

Die aggregierte allgemeine Bewertung zeigt eine negative Schiefe für NAO von -0,622 und für den Web-Chatbot von -0,528. Die Kurtosis liegt bei NAO bei -0,184 und bei Web-Chatbot -0,562, was insgesamt auf leicht zu hohen Werten verschobene und annähernd normalverteilte Antworten hinweist was in einem Fragebogen oft normal ist.

Die Analyse der Verteilungseigenschaften zeigt, dass die meisten Skalen bei beiden Systemen negative Schiefewerte aufweisen und damit tendenziell höhere Werte erreicht wurden. Die Kurtosis liegt überwiegend im Bereich leicht negativer Werte, was auf abgeflachte Verteilungen hinweist. Insgesamt bewegen sich die Verteilungen damit nahe am Normalbereich, mit einzelnen Abweichungen je nach Skala und System.

Fragenpaket	System	Schiefe	Kurtosis
LE	NAO	-0,647	-0,481
	Web-Chatbot	-0,752	0,337
AE	NAO	-0,248	-0,477
	Web-Chatbot	-0,795	-0,905
SE	NAO	-0,184	-0,375
	Web-Chatbot	0,128	-1,091
EB	NAO	-1,252	0,938
	Web-Chatbot	-0,662	-0,0262
VA	NAO	-0,493	-0,349
	Web-Chatbot	-0,809	-0,0318
Allgemein	NAO	-0,622	-0,184
	Web-Chatbot	-0,528	-0,562

Tabelle 25 - Schiefe und Kurtosis (UTAUT)

4.4.6 Überprüfung der Konsistenzreliabilität mittels Cronbach's Alpha

Die interne Konsistenz der erhobenen Skalen wurde anhand von Cronbach's Alpha überprüft, um die Zuverlässigkeit der eingesetzten Fragebögen zu evaluieren. Die Analyse wurde für die beiden untersuchten Systeme, den humanoiden Roboter NAO und den Web-Chatbot, durchgeführt.

Beim NAO zeigen sich durchweg hohe Reliabilitätswerte für die Skalen Leistungserwartung (0,908) und Verhaltensabsicht (0,912). Auch die Skala erleichternde Bedingungen weist mit 0,797 eine solide interne Konsistenz auf, während die Skala Anstrengungserwartung mit 0,457 deutlich niedrigere Werte erreicht. Besonders auffällig ist die Skala soziale Einflüsse, die beim NAO einen negativen Wert von -0,013 aufweist. Dieser Wert deutet darauf hin, dass die Items der Skala in diesem Kontext keine konsistente Messung der zugrunde liegenden Dimension darstellen. Der allgemeine Reliabilitätswert des UTAUT-Fragebogens für NAO beträgt 0,923 und zeigt damit, dass der Fragebogen insgesamt sehr zuverlässig ist, trotz einzelner Schwankungen auf Skalenebene.

Für den Web-Chatbot ergibt sich ein anderes Bild. Auch hier erreichen die Konstrukte Verhaltensabsicht (0,899) und Leistungserwartung (0,845) hohe Werte, die auf eine gute interne Konsistenz hinweisen. Die Skala Anstrengungserwartung weist mit 0,710 eine akzeptable bis gute Reliabilität auf, ebenso wie die Skala soziale Einflüsse mit 0,582. Die Skala erleichternde Bedingungen erreicht einen mittleren Wert von 0,560. Der allgemeine

Reliabilitätswert für den webbasierten Chatbot liegt bei 0,851, während der kombinierte Gesamtwert aller Skalen 0,891 beträgt. (vgl. Tabelle 26)

Die Reliabilitätsprüfung zeigt insgesamt hohe interne Konsistenzen der Skalen, insbesondere bei den Konstrukten Leistungserwartung und Verhaltensabsicht für beide Systeme. Auffällig ist der negative Wert beim Konstrukt „Soziale Einflüsse“ in der NAO-Gruppe sowie die geringeren Werte bei einzelnen Skalen des webbasierten Chatbots. Auf Gesamtfragebogenebene erreichen jedoch beide Systeme hohe Reliabilitätswerte.

Fragenpaket	**NAO**	**Web-Chatbot**
LE	0,908	0,845
AE	0,457	0,710
SE	-0,013	0,582
EB	0,797	0,560
VA	0,912	0,899
Allgemein	**0,923**	**0,851**
	0,891	

Tabelle 26 - Cronbach Alpha von NAO und Web-Chatbot (UTAUT)

4.5 Benutzerfreundlichkeit (SUS)

Die deskriptive Analyse des SUS-Wertes für den humanoiden Roboter NAO zeigt, dass die Antworten den gesamten Bereich der verwendeten Likert-Skala von 1,00 bis 5,00 abdecken. Der NAO Roboter erzielt in der SUS einen Mittelwert von 3,01. Mit einer Standardabweichung von 1,34 streuen die Einzelwerte jedoch stark, was auf heterogene Nutzererfahrungen schließen lässt (vgl. Tabelle 27)

Beim textbasierten Web-Chatbot fällt der SUS-Score etwas niedriger aus. Der Mittelwert liegt bei 2,78, während das Minimum auf 1,00 und das Maximum auf 5,00 erneut das volle Spektrum der Skala abbilden. Die Standardabweichung von1,43 ist leicht höher als beim NAO-Roboter. (vgl. Tabelle 28)

Die Auswertung der Benutzungsfreundlichkeit anhand von dem Geschlecht zeigt, dass der humanoide Roboter NAO in beiden Geschlechtergruppen als intuitiver wahrgenommen wird als der Web-Chatbot, wenn auch nur mit moderatem Abstand. Männer vergeben NAO einen Mittelwert von 3,08 und Frauen 2,90 Punkten, während die Werte für den Chatbot bei Männern 2,76 und bei Frauen bei 2,80 liegen. Damit fällt der geschlechtsspezifische Unterschied beim Roboter etwas deutlicher aus als beim textbasierten System. (vgl. Abbildung 27)

Der humanoide Roboter NAO wird von KI-Einsteigern als deutlich nutzerfreundlicher erlebt als der Web-Chatbot. Lernende mit wenig Erfahrung vergeben für NAO einen Mittelwert von 3,30,

der in der Gruppe mit mäßiger Vertrautheit auf 3,50 ansteigt. Sobald die Teilnehmenden jedoch als „vertraut“ eingestuft werden, sinkt der SUS-Wert auf 2,83. Der Web-Chatbot liegt in jeder Erfahrungsgruppe unter den NAO-Werten, zeigt aber einen gegenteiligen Trend. Für Lernende mit mäßiger Erfahrung beträgt der SUS-Score 2,60, bei „vertrauten“ steigt er auf 2,83 und hält sich bei „sehr vertrauten“ Anwender*innen mit 2,80 nahezu konstant. Damit nähert sich die Benutzerfreundlichkeit des Chatbots mit zunehmender Kompetenz dem Niveau des Roboters an, ohne es vollständig zu erreichen. (vgl. Abbildung 28)

Der humanoide Roboter NAO erreicht seinen höchsten SUS-Wert von 3,80 bei Lernenden, die KI nur gelegentlich einsetzen. Sobald der Roboter jedoch häufiger verwendet wird, sinkt der Score auf 2,98 und bei sehr häufiger Nutzung weiter auf 2,80. Beim Web-Chatbot liegen die SUS-Werte in einem engeren Spektrum und zeigen ein beinahe umgekehrtes Bild. Bei seltener Nutzung liegt der Score bei 2,70, fällt bei gelegentlichem Einsatz auf 2,60, steigt dann aber auf 2,88 in der Gruppe der häufigen Anwender*innen und bleibt mit 2,70 bei sehr intensiver Nutzung stabil. (vgl. Abbildung 29)

Frage	**min**	**$\bar{x}$**	**max**	**σ**
SUS1	2,00	3,70	5,00	0,82
SUS2	1,00	1,70	3,00	0,67
SUS3	4,00	4,20	5,00	0,42
SUS4	1,00	2,40	4,00	1,07
SUS5	3,00	3,90	5,00	0,57
SUS6	1,00	2,10	3,00	0,88
SUS7	2,00	4,30	5,00	0,95
SUS8	1,00	1,90	4,00	0,99
SUS9	3,00	3,90	5,00	0,99
SUS10	1,00	2,00	5,00	1,41
Allgemein	**1,00**	**3,01**	**5,00**	**1,34**

Tabelle 27 - Deskripitive Statistik - NAO (SUS)

Frage	min	$\bar{x}$	max	σ
SUS1	2,00	3,30	4,00	0,95
SUS2	1,00	1,30	2,00	0,48
SUS3	4,00	4,40	5,00	0,52
SUS4	1,00	1,30	2,00	0,48
SUS5	2,00	3,80	5,00	0,79
SUS6	1,00	2,50	5,00	1,27
SUS7	3,00	4,20	5,00	0,63
SUS8	1,00	1,80	5,00	1,32
SUS9	2,00	3,60	5,00	0,84
SUS10	1,00	1,60	4,00	0,97
Allgemein	**1,00**	**2,78**	**5,00**	**1,43**

Tabelle 28 - Deskripitive Statistik – Web-Chatbot (SUS)

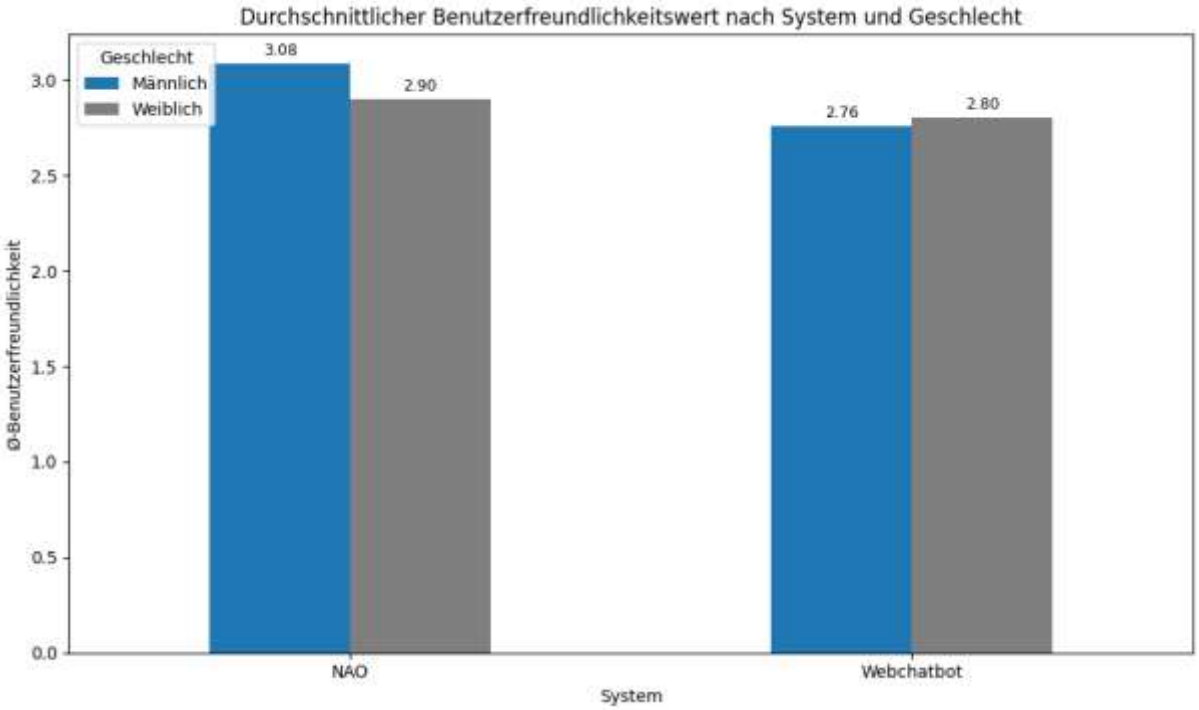

Abbildung 27 - Durchschnittlicher Benutzerfreundlichkeitswert nach System und Geschlecht

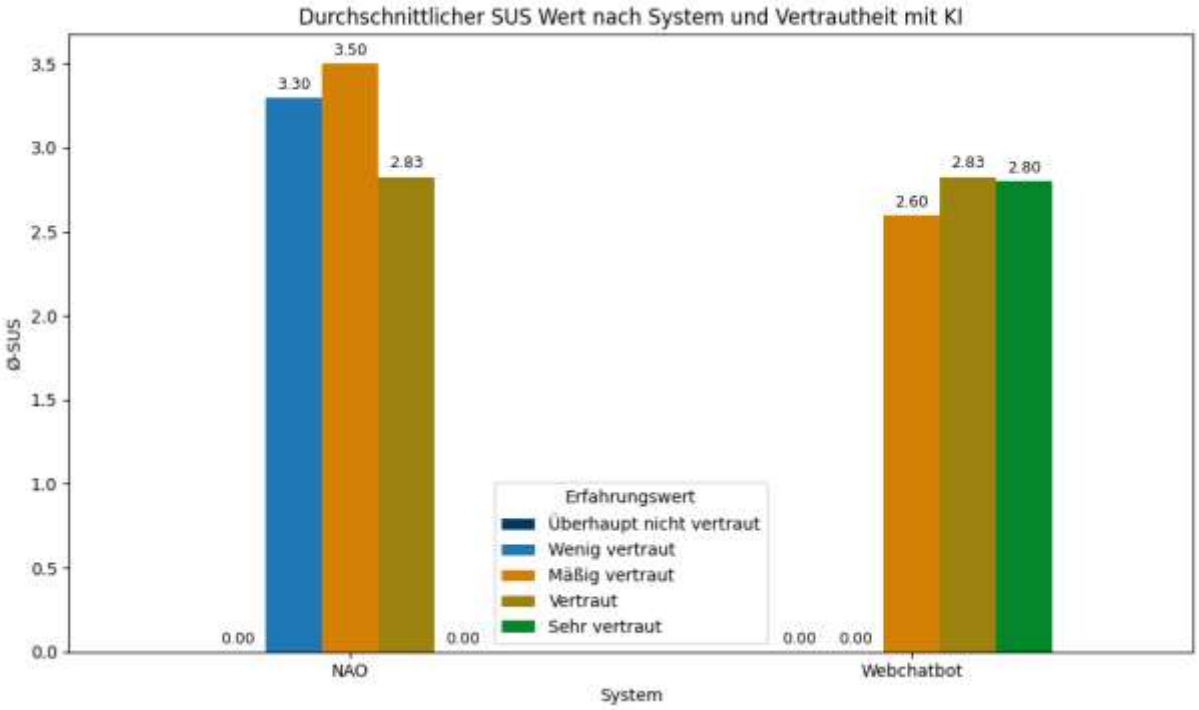

Abbildung 28 - Durchschnittlicher SUS Wert nach System und Vertrautheit mit KI

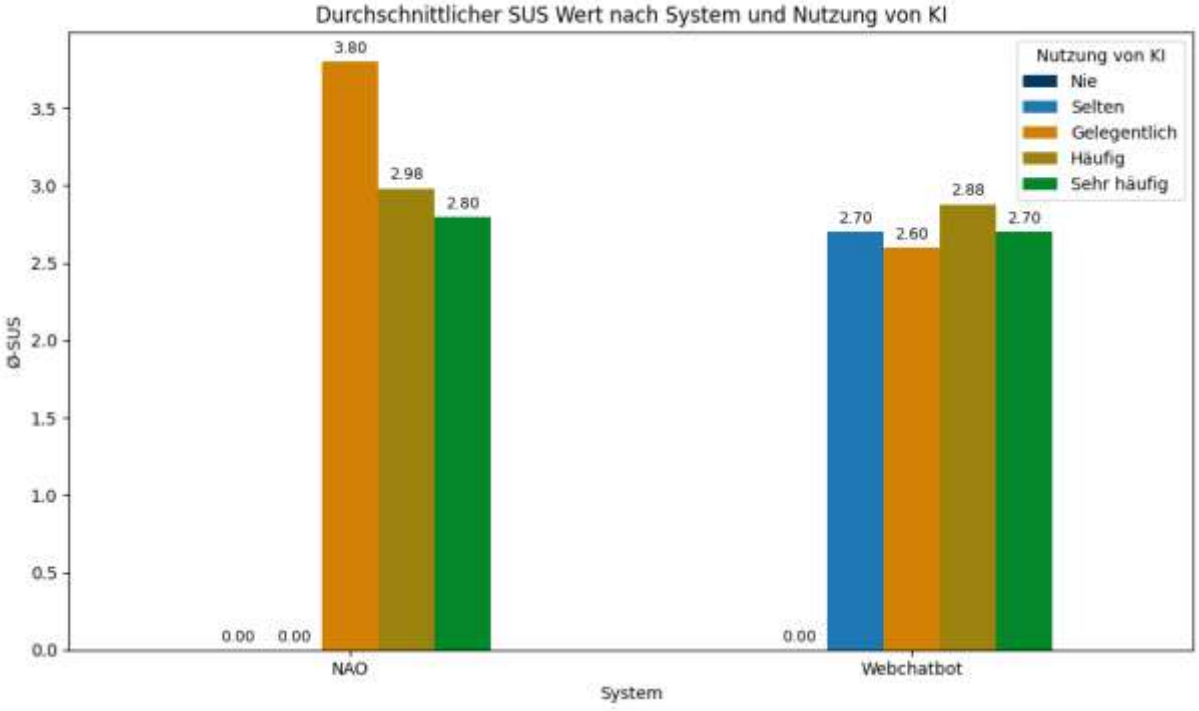

Abbildung 29 - Durchschnittlicher SUS Wert nach System und Nutzung von KI

4.5.1 Datenverteilung

Zur weiteren Beschreibung der Verteilungen der erhobenen Daten wurde die Schiefe (Skewness) und Kurtosis der SUS-Werte berechnet. Für den humanoiden Roboter NAO zeigt sich eine leicht negative Schiefe von -0,120, was darauf hinweist, dass die Antworten minimal zu höheren Werten hin verschoben sind. Die Kurtosis liegt bei -1,241, was auf eine leicht abgeflachte Verteilung im Vergleich zur Normalverteilung hinweist.

Beim Web-Chatbot fällt die Schiefe mit 0,039 nahezu neutral aus, sodass die Verteilung der Antworten symmetrisch ist. Die Kurtosis liegt bei -1,496, was ebenfalls auf eine flachere Verteilung hindeutet.

Für die aggregierte Gesamtbewertung über beide Systeme ergibt sich eine Schiefe von -0,048 und eine Kurtosis von -1,380. Dies deutet auf eine insgesamt nahezu symmetrische, leicht abgeflachte Verteilung hin. (vgl. Tabelle 29)

Die Ergebnisse zeigen, dass die Verteilungen der SUS-Werte bei beiden Systemen weitgehend symmetrisch ausfallen und durchgängig eine abgeflachte Form im Vergleich zur Normalverteilung aufweisen. Auch in der Gesamtbewertung über beide Systeme bleibt dieses Muster bestehen.

Fragenpaket	**System**	**Schiefe**	**Kurtosis**
SUS	**NAO**	-0,120	-1,241
	Web-Chatbot	0,039	-1,496
	Allgemein	-0,048	-1,380

Tabelle 29 - Schiefe und Kurtosis (SUS)

4.5.2 Überprüfung der Konsistenzreliabilität mittels Cronbach’s Alpha

Die interne Konsistenz des SUS-Fragebogens wurde mithilfe von Cronbach’s Alpha untersucht. Für den humanoiden Roboter NAO ergibt sich ein Wert von 0,425, während der webbasierte Chatbot einen negativen Wert von -2,428 aufweist. Dieser Wert liegen deutlich unter den allgemein als akzeptabel angesehenen Schwellenwerten und deuten auf eine unzureichende interne Konsistenz hin.

Die aggregierte Bewertung über beide Systeme zeigt einen Wert von 0,105, was ebenfalls auf eine nur sehr geringe Reliabilität des SUS-Fragebogens in der vorliegenden Stichprobe hinweist. (vgl. Tabelle 30)

Fragenpaket	NAO	Web-Chatbot
SUS	0,425	-2,428
Allgemein	0,105	

Tabelle 30 - Cronbach Alpha von NAO und Computer Agent (SUS)

5 Diskussion

Die vorliegende Diskussion ordnet die empirischen Befunde zum Vergleich eines humanoiden Roboters und eines webbasierten Chatbots auf Basis von Retrieval-Augmented Generation (RAG) im Bildungskontext ein und setzt sie in Beziehung zum aktuellen Forschungsstand. Darüber hinaus werden die methodischen und technischen Herausforderungen des Vorgehens kritisch beleuchtet und Limitationen transparent dargestellt. Abschließend wird ein Ausblick auf zukünftige Forschungsarbeiten gegeben.

5.1 Erkenntnisse der Umfrage

Die Teilnehmenden zeigen insgesamt eine hohe Vertrautheit mit digitalen Lernmitteln und KI, unabhängig von Geschlecht und Untersuchungsgruppe, und beide Geschlechter berichten sehr ähnliche Mittelwerte. Unterschiede liegen vor allem in der Spannweite der Selbsteinschätzung, bei Männern von mäßig vertraut bis sehr vertraut und bei Frauen von wenig vertraut bis vertraut, wobei Männer digitale Lernangebote etwas häufiger und gleichmäßiger nutzen. Beide Systeme weisen eine hohe Vertrautheit und praktische Nutzung auf, beim NAO treten vereinzelt sehr niedrige Werte auf, beim Web-Chatbot dagegen häufiger sehr hohe Werte.

Die Ergebnisse bestätigen eine insgesamt positive Benutzerfreundlichkeit, wobei NAO leichte Vorteile gegenüber dem Web-Chatbot zeigt, jedoch mit eingeschränkter Reliabilität (negativer Cronbach Alpha Wert). Unterschiede nach Vertrautheit und Nutzungshäufigkeit verdeutlichen, dass NAO besonders bei mittlerer Vertrautheit und gelegentlicher Nutzung punktet, während der Web-Chatbot bei regelmäßiger Anwendung überzeugt. Dies könnte damit zusammenhängen, dass die physische Präsenz und Interaktivität von NAO kurzfristig stärker motivieren, während der Web-Chatbot durch seine ständige Verfügbarkeit und vertraute Bedienbarkeit bei längerer Nutzung an Akzeptanz gewinnt. Daraus ergibt sich die Empfehlung, humanoide Roboter gezielt für punktuelle Lernszenarien einzusetzen, während Web-Chatbots für kontinuierliche Bildungsanwendungen geeigneter sind. Damit wird Hypothese 1 „Der NAO-Roboter wird von den Nutzer*innen als benutzerfreundlicher wahrgenommen als der webbasierte Chatbot.“ bestätigt. Dies stützt auch eine Studie von Barkana et al. [20], sie zeigten in einem experimentellen Setting mit dem NAO-Roboter plus einem erklärbaren Dialog-Manager eine klare Präferenz für die Roboter-Bedingung gegenüber einer rein webbasierten Variante, was auf eine bessere Benutzerfreundlichkeit hindeutet.

Betrachtet man die Leistungserwartung der beiden Systeme, zeigt sich dass der Web-Chatbot leicht im Vorteil ist und damit Hypothese 2 „Die Leistungserwartung ist beim NAO-Roboter höher ausgeprägt als beim Web-Chatbot.“ nicht bestätigt wird. Unterschiede nach Geschlecht fallen gering aus, während Vertrautheit und Nutzungshäufigkeit vor allem den Web-Chatbot begünstigen, der mit zunehmender Erfahrung stabil hohe Werte erreicht. Eine mögliche Erklärung liegt darin, dass der textbasierte Chatbot durch seine Verfügbarkeit und Routine-Nutzung als verlässlicher wahrgenommen wird, während humanoide Roboter durch ihre physische Präsenz zwar kurzfristig höhere Motivation und kognitive Beteiligung erzeugen können, aber stärker kontextabhängig bleiben. Die Studie von Venkatesh et al. [18] weist darauf hin, dass die Leistungserwartung bei Männern in der Regel höher ausfällt als bei Frauen. In dem vorliegenden Experiment konnte dieses Muster jedoch nicht bestätigt werden. Da der Leistungserwartungsdurschnitt des Web-Chatbot‘s etwas höher liegt, wird Hypothese 2 nicht bestätigt. Trotz fehlender Bestätigung dieser Hypothese, zeigen Wedenborn et al. [21], dass Lernende mit einem physischen Roboter signifikant bessere Erinnerungsleistungen erzielen als mit einer rein digitalen Alternative, was auf eine höhere tatsächliche Lerneffektivität hindeutet und somit Leistungserwartung positiv beeinflussen könnte. Auch Fung et al. [14] berichten von einem starken Anstieg kognitiver Beteiligung, intrinsischer Motivation und Lernerfolg, wenn humanoide Roboter als Lehrassistenten eingesetzt werden. Diese Ergebnisse sprechen für das Potenzial humanoider Systeme zur Effizienzsteigerung beim Lernen.

Die Ergebnisse weisen auf eine moderate Aufwandserwartung hin, wobei der Web-Chatbot einen leichten Vorteil gegenüber dem NAO erzielt und zugleich von zunehmender Erfahrung und häufiger Nutzung profitiert. Die Reliabilität ist beim Web-Chatbot akzeptabel, während die Werte beim NAO weniger belastbar erscheinen, was die Interpretation einschränkt. Die Analyse deutet darauf hin, dass die Erfahrung der Nutzer*innen tatsächlich einen moderierenden Einfluss auf die Aufwandserwartung haben könnte, die Venkatesh et al. [18] identifiziert haben. Allerdings lässt sich dies aufgrund der geringen Stichprobengröße und möglicher Unschärfen nicht eindeutig bestätigen. Die Ergebnisse zeigen auch, dass die durchschnittliche Aufwandserwartung bei männlichen Teilnehmenden höher ausfällt als bei weiblichen, sowohl im Fall von NAO als auch beim textbasierten Chatbot. Damit widerspricht das Ergebnis der Annahme aus dem UTAUT-Modell, wonach Aufwandserwartung insbesondere für Frauen eine größere Rolle spielt [18]. Auffällig ist, dass NAO vor allem bei gelegentlicher Nutzung höhere Werte erreicht, während der Web-Chatbot seine Stärken bei sehr häufiger Anwendung ausspielt. Humanoide Roboter können durch ihre Präsenz und Interaktivität zwar motivierend wirken, ihre Bedienung wird aber oft als aufwändiger empfunden, weil sie komplexer und weniger flexibel sind. Web-Chatbots dagegen wirken vertrauter, da viele Nutzer*innen den Umgang mit textbasierten Systemen gewohnt sind, dadurch lassen sie sich besonders bei längerer und regelmäßiger Nutzung leichter handhaben. Hypothese 3 „Die Aufwandserwartung ist beim Web-Chatbot höher ausgeprägter als beim NAO-Roboter“ wird bestätigt. Es gibt Hinweise, dass verkörperte Agenten nicht zwingend effizienter oder leichter zu bedienen sind. Lukasik et al. [22] bestätigen auch in der Studie „From robots to chatbots: unveiling the dynamics of human-AI interaction“, dass physisch verkörperte Agenten wie humanoide Roboter emotionaler und sozial ansprechender wahrgenommen werden, aber nicht notwendigerweise effizienter oder leichter zu bedienen sind.

Die Analyse zeigt einen mittelhohen sozialen Einfluss in beiden Systemgruppen, wobei der Web-Chatbot in mehreren Subgruppen gleichauf oder sogar höher bewertet wird, sodass Hypothese 4 nicht bestätigt werden konnte. Während die Messung beim NAO durch geringe Reliabilität eingeschränkt ist, weist der Web-Chatbot eine konsistente Skala auf. Eine mögliche Erklärung hierfür liegt darin, dass textbasierte Systeme stärker im Alltag verankert sind und ihre Nutzung durch das soziale Umfeld, etwa durch den verbreiteten Einsatz von Chatbots in Service und Kommunikation, eher als selbstverständlich wahrgenommen wird. Humanoide Roboter dagegen sind in Bildungskontexten noch selten und werden daher weniger stark durch Peers oder Lehrkräfte beeinflusst, sondern eher durch die Situation der Interaktion. Studien verorten den Einfluss stärker bei Lehrpersonen, Peers und Kontexten sowie bei Moderatoren wie Geschlecht und Erfahrung [224], [225].

Die erleichternden Bedingungen werden sowohl durch die Vertrautheit mit dem System als auch durch das Geschlecht beeinflusst. Beim NAO steigen die Bewertungen mit wachsender Vertrautheit kontinuierlich an und erreichen ihr Maximum bei mittlerer Vertrautheit, während der Web-Chatbot bereits auf hohem Niveau stabil bleibt und von häufiger Nutzung besonders profitiert. Geschlechterunterschiede zeigen ein differenziertes Muster, Frauen berichten beim NAO deutlich höhere Werte als Männer, was darauf hindeuten könnte, dass die physische Präsenz und Interaktivität des Roboters für sie stärker als unterstützend wahrgenommen werden. Beim Web-Chatbot kehrt sich das Muster leicht um, da Männer ihn als geringfügig einfacher nutzbar einschätzen, vermutlich aufgrund der Vertrautheit mit textbasierten Interfaces aus alltäglichen Anwendungen. Somit wird die Hypothese 5 „Die erleichternden Bedingungen wirken beim NAO stärker als beim Web-Chatbot.“ nicht bestätigt, da die zusammengeführten Muster aus zentraler Tendenz, Verteilungsform und Subgruppenanalysen insgesamt zugunsten des Web-Chatbots sprechen. Durak et al. [225] bestätigen, dass Web-Chatbot‘s stark von stabil wahrgenommenen Unterstützungsbedingungen profitiert, was in dieser Stichprobe eines leicht besseren Durchschnittswerts und konsistenter Verteilung in dieser Dimension stützt. Obwohl die Nutzungsdauer in der vorliegenden Untersuchung nicht erfasst wurde, betonen Durak et al. [225] in der Studie „Predicting the use of chatbot systems in education“ die zentrale Rolle der Nutzungsdauer als signifikanter Einflussfaktor auf die erleichternden Bedingungen. Insgesamt zeigt sich, dass erleichternde Bedingungen nicht nur systemspezifisch, sondern auch stark von individuellen Faktoren abhängen. Web-Chatbots profitieren dabei von einer allgemeinen Akzeptanz und bekannten Nutzungsmustern, während humanoide Roboter gezielt bei Nutzenden und in Kontexten eingesetzt werden sollten, in denen ihre besondere Form der Interaktion Mehrwert stiftet.

Die Resultate verdeutlichen, dass beide Systeme insgesamt positiv bewertet werden und die Verhaltensabsicht in beiden Fällen über dem Skalenmittelpunkt liegt. Zwar weist der Web-Chatbot einen kleinen Vorsprung auf, doch die Differenzen bleiben insgesamt gering und Hypothese 6 „Die Verhaltensabsicht zur weiteren Nutzung ist beim NAO-Roboter höher als beim Web-Chatbot.“ wird nicht bestätigt. Auffällig ist, dass NAO insbesondere bei gelegentlicher Nutzung eine hohe Intention hervorruft, während der textbasierte Chatbot unabhängig von der Nutzungshäufigkeit stabil bleibt und mit steigender Erfahrung sogar an Zustimmung gewinnt. Geschlechtsspezifische Unterschiede zeichnen ein differenziertes Bild, Männer zeigen eine deutlich höhere Tendenz für den Web-Chatbot, während Frauen NAO etwas stärker bevorzugen. Eine mögliche Erklärung hierfür liegt darin, dass humanoide Roboter bei Interaktion durch ihre physische Präsenz kurzfristig stärkeres Interesse wecken,

während der Web-Chatbot durch seine Vertrautheit und Alltagstauglichkeit eine konsistente und nachhaltige Nutzungsintention fördert. Damit wird deutlich, dass sich beide Systeme in ihren Stärken ergänzen und ihre Akzeptanz stark von Kontext und Nutzermerkmalen abhängt. Die Literatur zeigt, dass die Verhaltensabsicht bei der Nutzung von Chatbots signifikant durch deren einfache Bedienbarkeit und hohe Leistungserwartung steigt [23]. Diese positiven Effekte bleiben auch über längere Zeiträume hinweg stabil. Im Gegensatz dazu zeigen humanoide Roboter seltener einen vergleichbaren, nachhaltigen Anstieg der Nutzungsmotivation [23]. Es zeigte sich, dass erleichternde Bedingungen und sozialer Einfluss einen signifikanten Einfluss auf die Verhaltensabsicht ausüben, insbesondere bei zunehmender Vertrautheit mit der Technologie [225].

Das vollständige UTAUT-Konstrukt zeigt für beide Systeme eine insgesamt positive Grundhaltung, mit einem kleinen, aber konsistenten Vorteil für den Web-Chatbot. Während NAO vor allem in frühen Phasen durch seine physische Präsenz und Neuheit überzeugt, gewinnt der Web-Chatbot mit zunehmender Erfahrung und häufiger Nutzung an Stabilität und Akzeptanz. Geschlechterunterschiede bleiben gering, zeigen aber, dass Frauen beim NAO und Männer beim Web-Chatbot leicht höhere Werte verzeichneten. Die stärker linksorientierte Verteilung beim NAO verdeutlicht sehr hohe Zustimmungen einzelner Teilgruppen, die den Gesamtdurchschnitt abflachen, während die Urteile beim Web-Chatbot stabil um ein hohes Zentrum gruppiert sind. Dieses Muster erklärt, warum der Web-Chatbot trotz kleiner Vorteile in Mittelwerten praxisrelevant besser abschneidet, da er weniger kontextabhängig ist. Eine mögliche Begründung liegt auch darin, dass textbasierte Chatbots, verstärkt durch den Erfolg von OpenAI und vergleichbaren Plattformen, bereits stärker in den Alltag integriert sind und somit von einer größeren Vertrautheit profitieren. Dadurch erscheinen sie leichter zugänglich, während humanoide Roboter zusätzliche Infrastruktur und Einarbeitung erfordern. Didaktisch ergibt sich ein komplementäres Einsatzfeld, der Roboter eignet sich als motivierender Türöffner und für soziale Einbindung, der Chatbot für Routinen und Handhabbarkeit. Insgesamt bestätigt sich, dass die kleinen Vorteile des textbasierten Chatbots nicht nur statistisch, sondern auch durch technologische Trends und Marktverbreitung gestützt sind. Die Hypothese 7 „Der NAO erzielt höhere Akzeptanzwerte als der Web Chatbot“ wird nicht bestätigt. In einer Studie von Tarlan et al. [24] wurde der humanoide Roboter zwar als sympathischer und geringfügig intelligenter wahrgenommen, doch in der grundlegenden Akzeptanz war kein klarer Vorteil zum Webchatbot erkennbar. Fragakis et al. [226] untersuchten in ihrer Studie den Lernmitteleinsatz durch den NAO-Roboter im Vergleich zu textbasierten Interaktionssystemen wie LLMs und Suchmaschinen. Sie betonen, dass der NAO nicht nur durch seine physische Präsenz und soziale Interaktion motiviert, sondern auch adaptive und einfühlsame Rückmeldungen bietet, die Lernende besonders unterstützen, insbesondere bei komplexen Aufgaben wie der Wissensvermittlung [226]. Ein weiterer wichtiger Punkt wurde von Koyuturk et al. [145] erforscht, dass klare Prompting Richtlinien die Benutzerfreundlichkeit und die Antwortqualität bei LLM gestützten Chatbots erhöhen, besonders bei geringerer Erfahrung in der Anfrageformulierung.

Die Ergebnisse der Lernmessung zeigen, dass NAO im Pretest einen höheren Ausgangswert aufweist und sich im Verlauf verschlechtert, während der Web-Chatbot auf einem weitgehend konstanten Niveau bleibt. Im Posttest konvergieren beide Systeme auf nahezu identische Mittelwerte, sodass Hypothese 8 „Der Einsatz des NAO führt zu höheren Wissensvermittlung als der Einsatz eines webbasierten Chatbots.“ nicht bestätigt wird. Dies deutet darauf hin, dass der humanoide Roboter in diesem Experiment keinen stärkeren Lernfortschritt anstoßen

konnte, während Chatbots ein stabiles, aber weniger dynamisches Lernmuster zeigen. Eine plausible Erklärung hätte darin liegen können, dass humanoide Roboter durch ihre soziale und körperliche Verkörperung stärkere Aufmerksamkeit binden und Lernende aktivieren, wodurch gerade in frühen Lernphasen ein zusätzlicher Motivationsschub entsteht. Der webbasierte Chatbot hingegen bietet konstante, quellengestützte Unterstützung, ohne dieselben Aktivierungseffekte auszulösen, wodurch er vor allem in stabilen Lernroutinen seine Stärken zeigt. Eine Studie von Wedenborn et al. [21] hätte diese Beobachtung bestätigen können, dort wurde der Wortabruf nach Vokabeltraining erheblich verbessert, wenn dieses mit einem physischen Roboterkopf stattfand, im Vergleich zu einem Avatar oder einer rein auditiven Präsentation. Fung et al. [14] liefern hierfür auch eine fundierte Erklärung. Der humanoide Roboter sichert durch emotional-kognitive Aktivierung und gesteigerte Engagementwerte einen beträchtlichen Lernfortschritt, insbesondere dort, wo Motivation und Teilnahme entscheidend sind [14]. Eine mögliche Erklärung dafür, dass die Hypothese nicht bestätigt werden konnte, wäre eine Verzerrung durch mangelnde Motivation der Studierenden.

Die Forschungsfrage dieser Arbeit lautet: Wie beeinflusst die Verwendung eines humanoiden Roboters im Vergleich zu einem Web-Chatbot mit RAG die Akzeptanz und die Benutzerfreundlichkeit im Bildungsbereich? Die Ergebnisse zeigen, dass beide Systeme insgesamt positiv bewertet werden, jedoch mit unterschiedlichen Stärken. Der humanoide Roboter NAO entfaltet insbesondere in frühen Lernphasen durch seine physische Präsenz und soziale Interaktion eine motivierende Wirkung, die sich in gesteigertem Engagement und messbaren Lernfortschritten widerspiegelt. Der Web-Chatbot hingegen profitiert von seiner Alltagsnähe, Stabilität und einfachen Handhabung, wodurch er bei längerer Nutzung konsistent akzeptiert wird und in der Benutzerfreundlichkeit leicht im Vorteil ist. Ein deutlicher Unterschied zwischen beiden Technologien lässt sich nicht feststellen, was auch darauf zurückzuführen ist, dass Chatbots durch die starke Verbreitung von LLM-gestützten Anwendungen bereits fest in den Alltag vieler Nutzer*innen integriert sind.

Ein ergänzender Aspekt in diesem Zusammenhang ist das Konzept des Uncanny Valley. Humanoide Roboter wie der NAO bewegen sich in einem Wahrnehmungsspektrum, das starke Ähnlichkeit zum Menschen nahelegt, jedoch nicht vollständig erreicht. Diese fast-menschliche Erscheinung kann bei Nutzer*innen Irritationen oder ein ambivalentes Gefühl auslösen insbesondere dann, wenn Erwartung und tatsächliches Verhalten nicht übereinstimmen. Die Studie von Kim et al. [227] bestätigt, dass humanoide Roboter durch ihre physische Präsenz und LLM-Unterstützung zu einer stärkeren Anthropomorphisierung führen, was die Interaktion intuitiver, aber auch weniger zielgerichtet macht. Im Gegensatz dazu werden Web-Chatbots nüchterner und funktionaler genutzt, mit höherer Prompt-Präzision. Das Uncanny Valley kann somit ein Erklärungsansatz dafür sein, warum humanoide Roboter trotz hoher Sympathiewerte nicht automatisch zu höherer Akzeptanz führen, insbesondere bei intensiverer oder kritischer Nutzung. So wird ihnen verstärkt eine sozial-emotionale Kompetenz zugeschrieben, was dazu führen kann, dass Interaktionen stärker anthropomorphisiert werden. Während Web-Chatbots tendenziell nüchtern und funktional genutzt werden mit höherer Prompt-Präzision besteht bei humanoiden Robotern eher die Tendenz, weniger explizit zu fragen und intuitiver zu interagieren, da die Nutzer*innen annehmen, mit einem "sozialen Akteur" zu kommunizieren. [227]

Für Bildungseinrichtungen ergibt sich aus dieser Forschungsarbeit die Empfehlung, humanoide Roboter gezielt in Szenarien einzusetzen, in denen Motivation, Aktivierung und soziale Einbindung von Lernenden im Vordergrund stehen. Web-Chatbots hingegen sind

besonders geeignet für kontinuierliche Übungs- und Unterstützungsphasen, in denen Routine, Verfügbarkeit und eine einfache Integration in den Lernalltag entscheidend sind. Das größte Potenzial liegt daher in einem Zusammenspiel beider Technologien, der Roboter kann als motivierender Türöffner und sozialer Verstärker wirken, während der Chatbot als verlässlicher Begleiter den Lernprozess langfristig unterstützt. Wichtig ist dabei, dass Bildungseinrichtungen die jeweiligen Einsatzszenarien klar definieren, um die Stärken der Technologien optimal zur Geltung zu bringen. Zudem sollte bedacht werden, dass die Akzeptanz stark von der Vertrautheit der Lernenden abhängt, weshalb eine schrittweise Einführung sinnvoll erscheint. Lehrkräfte können dabei eine zentrale Rolle spielen, indem sie beide Systeme in didaktische Konzepte einbetten und deren Nutzung pädagogisch begleiten. Auch eine kontinuierliche Evaluation der Wirksamkeit ist empfehlenswert, um Erfahrungen aus der Praxis in die Weiterentwicklung der Systeme zurückfließen zu lassen. Ein weiterer Tipp ist, Lernende aktiv in die Gestaltung und Reflexion der Nutzung einzubeziehen, um Selbstwirksamkeit und Partizipation zu fördern. Insgesamt zeigt sich, dass nicht eine einzelne Technologie die Lösung darstellt, sondern eine bewusste Kombination beider Ansätze den größten Mehrwert für den Bildungsbereich bietet. Damit leistet der reflektierte Einsatz von KI-Technologien wie humanoiden Robotern und Chatbots einen wertvollen Beitrag zur Erreichung des Sustainable Development Goals 4, inklusive gerechter und hochwertiger Bildung für alle. Durch ihre differenzierten Einsatzmöglichkeiten können sie Bildungszugänge verbessern, Lernprozesse individualisieren und zur Chancengleichheit beitragen. Entscheidend ist dabei, dass technologische Innovation stets mit pädagogischer Verantwortung und nachhaltiger Zielorientierung einhergeht.

5.2 Herausforderungen und Limitierungen

Diese wissenschaftliche Untersuchung weist mehrere Beschränkungen auf. Die Beschränkungen können die Interpretation der Forschungsergebnisse beeinflussen. In diesem Kapitel wird ein detaillierter Überblick über die Beschränkungen dieser Studie gegeben und ihre möglichen Auswirkungen auf die Aussagekraft der Ergebnisse diskutiert.

5.2.1 Technische Herausforderungen

Dieses Kapitel bündelt die technischen Herausfordernden, die den Einsatz der Stichprobe prägen. Während der Web-Chatbot und der humanoide Roboter die Interaktion sichtbar machen, entstehen Verständnis, Quellenwahl und Antwortbildung im in der RAG-Pipeline. Entsprechend bestimmen nicht nur die Frontend-Interfaces, sondern vor allem die Infrastruktur dahinter die wahrgenommene Qualität.

5.2.1.1 Abhängigkeiten von LLM’s

Während der Stichprobenerhebung zu einem ungeplanten Ausfall des LLM-Anbieters (OpenAI) von rund 20 Minuten. In diesem Zeitraum war keine stabile Antwortgenerierung möglich, Anfragen liefen in Timeouts oder kehrten mit Fehlern zurück. Die betroffene Interaktion wurde aus der Stichprobe entfernt, um Verzerrungen in den Messwerten zu vermeiden. IT-Störungen stellen externe Einflüsse dar, die nicht der eigentlichen Systemleistung (z. B. Retrieval-Qualität oder Prompting) zuzuschreiben sind.

5.2.1.2 NAO Roboter

Die Leistungsfähigkeit des NAO wird durch mehrere technischen Faktoren begrenzt, die sich unmittelbar auf Planung und Aussagekraft von Experimenten auswirken. Die Audioseite ist sehr anfällig, Geräusch- und Spracherkennung reagieren in halligen Räumen oder bei Störgeräuschen fehlerhaft, was Dialoge unter realen Bedingungen erschwert. Auf der Softwareseite erschwert das veraltete NAOqi-Ökosystem die Nutzung moderner Bibliotheken. Schließlich wirkt die Verbindung als weitere Herausforderung, die Abhängigkeit vom WLAN kann zu Verzögerungen, Dropouts und damit zu inkonsistentem Verhalten führen. Aufgrund der schlechten Audioqualität des integrierten NAO-Mikrofons wurde das Laptop-Mikrofon verwendet.

5.2.1.3 Spracherkennung

Menschen verwenden nicht nur eine Sprache, sondern zahlreiche Varianten wie regionale Dialekte. Für automatische Spracherkennungssysteme bedeutet dies, dass identische Inhalte in Bezug auf Lautstärke, Wortschatz und Grammatik in sehr unterschiedlicher Form realisiert werden können.

Im Fall dieser wissenschaftlichen Arbeit hat die Erkennung nicht immer zuverlässig funktioniert, weil ein whisper Modell von OpenAI verwendet wurde. Solche kompakten Modelle sind zwar auf geringe Latenz, Offline-Betrieb und Datenschutz optimiert, verfügen jedoch meist über weniger Parameter und kleinere Lexika bzw. Sprachmodelle, bieten eine geringere Abdeckung für Dialekte, Akzente und Code-Switching und sind durch stärkere Kompression/Quantisierung gegenüber Rauschen insgesamt weniger robust.

5.2.1.4 RAG

Dieses Kapitel skizziert die zentralen Herausforderungen der RAG Pipeline. Auch wenn die Interaktion sichtbar über Web-Chatbot und humanoiden Roboter (NAO) erfolgt, liegt die „Intelligenz“ im Hintergrund. Der Chatbot und der Roboter verkörpern diese Intelligenz, sie sind Interface und Ausdruckskanal, nicht der Ort, an dem Verständnis, Quellenwahl und Antwortbildung entstehen. Entsprechend konzentrieren sich die wichtigsten Herausforderungen weniger auf die Oberfläche der Interaktion als auf die Robustheit, Verlässlichkeit und Latenz des „Gehirns“ im Backend.

Im Praxisbetrieb zeigte sich, dass die technische Architektur der Hauptfaktor für die Antwortzeit ist. Die Kette NAOqi (Python 2.7) → Bridge → Docker-Backend erzeugt zusätzliche Übergänge zwischen Prozessen und über das Netzwerk, wodurch Daten mehrfach ver- und entpackt sowie hin- und hergeschickt werden müssen, diese Schritte verursachen den größten Teil der Verzögerung über alle Phasen der Pipeline (Query-Kondensation, Retrieval, Re-Ranking, Generierung). Ein gewisser Gesprächskontext bleibt zwar notwendig, damit Rückfragen (z. B. „Was ist Computer Vision?“ oder „Bitte genauer.“) korrekt verankert werden, im Vergleich zu den architekturbedingten Kosten fällt der zusätzliche Token-Umfang aus dem Chatverlauf jedoch kaum ins Gewicht.

5.2.1.5 Prompt Engineering

Einen System-Prompt so zu gestalten, dass ein LLM konsequent genau im gewünschten Stil, Format und inhaltlichen Rahmen antwortet, ist schwierig. Sprachmodelle sind probabilistische Systeme, sie reagieren sensibel auf Wortwahl, Reihenfolge von Anweisungen, Beispiele und Kontextlänge. Schon kleine Änderungen können Wirkung und Zuverlässigkeit messbar verschieben.

5.2.2 Limitationen der Methode

Dieses Kapitel erläutert die Grenzen der eingesetzten Methodik und ordnet die Befunde hinsichtlich Validität und Übertragbarkeit ein.

5.2.2.1 Fragebogen

Ein wichtiger Kritikpunkt betrifft den eingesetzten Fragebogen. Zwar wurden alle relevanten UTAUT-Konstrukte erfasst, dennoch ist das Instrument inhaltlich begrenzt und schränkt die Aussagekraft teilweise ein. So wurde zum Beispiel nicht erhoben, wie lange die Teilnehmenden tatsächlich mit dem jeweiligen System interagiert haben. Diese Angabe hätte Hinweise auf Nutzungserfahrung, mögliche Desinteresse oder Ersteindruck-Effekte liefern können und wäre gerade für die Bewertung von Aufwand und Leistungserwartung hilfreich gewesen. Auch die sogenannte Prompt-Kompetenz, beziehungsweise der Umgang mit Eingaben im Chatbot, wurde nicht erfasst, obwohl sie nachweislich die Nutzungserfahrung beeinflussen kann. Zudem wurde der Fragebogen für beide Systeme identisch eingesetzt, ohne stärker auf deren Besonderheiten einzugehen (z. B. physische Präsenz beim NAO oder reines Text-Feedback beim Chatbot). Das erhöht zwar die Vergleichbarkeit, verzerrt aber systemspezifische Unterschiede.

5.2.2.2 Reliabilität der Konstrukte

In der vorliegenden Untersuchung zeigte sich für die NAO-Gruppe für den Sozialen Einfluss jedoch ein negativer Cronbachs Alpha von -0,013. Auch beim Web-Chatbot weist das Konstrukt „System Usability Scale (SUS)“ eine unzureichende Reliabilität auf. Damit ist ebenso fraglich, ob die erhobenen Items die zugrunde liegende Dimension stabil erfassen. Diese Werte weisen darauf hin, dass die Items der Konstrukte nicht konsistent miteinander variieren und somit keine interne Konsistenz vorliegt. Theoretisch bedeutet dies, dass die erhobene Skala in dieser Gruppe nicht zuverlässig die intendierte Dimension abbildet.

Trotz dieser unzureichenden Reliabilität werden die Konstrukte SE der NAO-Stichprobengruppe und SUS der Web-Chatbot-Gruppe in den Analysen berücksichtigt. Die Entscheidung hierfür wurde getroffen, um die Vergleichbarkeit zwischen den Untersuchungsgruppen zu gewährleisten und den theoretisch vorgesehenen Aufbau des Forschungsdesigns nicht zu verzerren.

Es muss dennoch betont werden, dass die Interpretation der Ergebnisse für dieses Konstrukt in der NAO-Gruppe mit Vorsicht zu erfolgen hat.

5.3 Ausblick für zukünftige Forschung

Künftige Arbeiten sollten das Studiendesign in der Breite und Tiefe ausbauen. Empfohlen ist eine Skalierung der Stichprobe auf etwa 40-300 Teilnehmende, verteilt über mehrere Studiengänge und Standorte, um externe Validität und Generalisierbarkeit zu erhöhen. Zusätzlich sollte die Perspektive der Dozent*innen systematisch einbezogen werden, etwa durch parallele Befragungen oder Interviews um Kontextfaktoren der Lehre miterfassen. Datenschutz- und Vertrauensfragen verdienen dabei besondere Aufmerksamkeit. Neben einer transparenten Einwilligung und Datenminimierung könnten „Datenschutzbedenken“ und „Vertrauen“ als eigene Konstrukte erhoben und als Moderatoren im Modell geprüft werden. Darüber hinaus sollte die Nutzungsdauer mitaufgenommen werden.

Methodisch bietet sich eine stärkere statistische Fundierung an. Vorab durchgeführte Power-Analysen können die Fallzahlplanung absichern. Für das UTAUT-Modell empfiehlt sich eine Faktorenanalyse mit anschließender Strukturgleichungsmodellierung, inklusive Tests auf Messinvarianz zwischen Studiengängen und Rollen (Studierende vs. Lehrende). Der SUS-Wert sollte hinsichtlich Reliabilität und möglicher Skalenverzerrungen geprüft werden. Für den Pre- und Post-Wissenstest sind Varianzanalysen mit Messwiederholung oder gemischte Linear-Modelle sinnvoll, ergänzt um Effektstärken, Konfidenz- und ggf. Bayes-Intervalle.

Auf technischer Ebene ergeben sich mehrere konkrete Ansatzpunkte für zukünftige Weiterentwicklungen. So könnte die Retrieval-Komponente durch eine Kombination klassischer Verfahren (BM25) mit modernen Embedding-Ansätzen ergänzt werden, um die Trefferqualität zu erhöhen. Darüber hinaus bietet sich ein semantisches Chunking mit Metadaten-Annotation an, um kontextgetreuere Antworten zu ermöglichen. Um Halluzinationen der KI zu reduzieren, wäre die Integration von Kontrollmechanismen sinnvoll, beispielsweise durch systematisches Zitieren der verwendeten Quellen sowie durch Faithfulness-Checks. Auch die Qualität der Antwortgenerierung könnte schrittweise gesteigert werden, etwa durch Prompt-Tuning oder Adapter-Tuning. Schließlich sollte bei gesprochener Interaktion auf eine robustere Spracherkennung gesetzt werden, die auch Dialektvarianten abdeckt, ergänzt durch eine serverseitige Fallback-Lösung zur Absicherung der Verarbeitung.

6 Literaturrecherche – PRISMA 2020

Dieses Kapitel beschreibt die systematische Literaturrecherche, die die theoretische Grundlage der Arbeit legt und den aktuellen Forschungsstand strukturiert. Ausgehend von präzisen Fragestellungen wurden Suchstrategien, Datenbanken sowie Ein- und Ausschlusskriterien definiert, um relevante und qualitativ hochwertige Quellen zu identifizieren.

6.1 Prozess der Literaturrecherche

Der Prozess der Literaturrecherche wurde systematisch nach dem PRISMA 2020 Flow-Diagramm gestaltet, um eine strukturierte und transparente Auswahl relevanter wissenschaftlicher Publikationen zu gewährleisten. Ziel war es, eine präzise und fundierte Grundlage für die Forschungsarbeit zu schaffen.

Zunächst wurde ein breiter Überblick über das Themenfeld gewonnen. Dazu wurden allgemeine Suchbegriffe verwendet, um eine möglichst umfassende Sammlung potenziell relevanter Publikationen zu erhalten. Im Anschluss daran wurden die Suchanfragen sukzessive präzisiert, um gezielt wissenschaftliche Arbeiten zu identifizieren, die einen spezifischen Bezug zur Forschungsfrage aufweisen.

Im nächsten Schritt wurden die Publikationen zunächst anhand des Titels bewertet. Passte der Titel zum Thema, wurde die Publikation in einem Excel-Dokument erfasst, das als zentrale Datenbank für die weitere Analyse diente. Anschließend wurde das Abstrakt oder die Einleitung jeder Publikation gründlich gelesen. Publikationen, die anhand des Abstracts als irrelevant eingestuft wurden, wurden aus dem weiteren Screening ausgeschlossen. War der Abstract jedoch vielversprechend und inhaltlich passend, wurde die Publikation vollständig gelesen und in die finale Analyse aufgenommen.

Dieser iterative und strukturierte Prozess ermöglichte es, die relevanten wissenschaftlichen Arbeiten gezielt auszuwählen und dabei eine Balance zwischen Breite und Tiefe der Recherche sicherzustellen.

6.2 Programme und Datenbanken

Für das Tracking und die Organisation der Ergebnisse der Literaturrecherche wurden die Tools Mendeley und Excel verwendet. Diese Programme ermöglichten eine strukturierte Verwaltung und effiziente Auswertung der gesammelten Quellen, die durch die Literaturrecherche gefunden wurden.

Die Literaturrecherche wurde mithilfe renommierter wissenschaftlicher Datenbanken durchgeführt, darunter Google Scholar, ACM Digital Library, IEEE Xplore, SpringerLink, Arxiv, Taylor & Francis, SpringerOpen, Nature, MDPI, Frontsier, ScienceDirect und Google Suche. Diese Datenbanken gewährleisteten den Zugriff auf qualitativ hochwertige und aktuelle Forschungsergebnisse, die die Grundlage der Diplomarbeit bilden.

6.3 Suchbegriffe

In den verschiedenen Forschungsdatenbanken wurden die folgenden Suchbegriffe verwendet und Filter verwendet.

6.3.1 ACM

Suchbegriff	**Filter**	**Anzahl**
Title: empathy robot	/	44
Title: SUS	/	23
Title: retrieval-augmented generation	/	137
Title: emotion robot	/	217
Title: NAO	/	46
Title: chatbot	/	639
Title: theory of mind	/	57
Title: Uncanny Valley	/	27
Title: user acceptance	/	376
Title: technology acceptance	/	346
Title: anthropomorphism	/	88
Title: user experience impact	/	80
Title: RAG Design	/	8
Title: humanoid robots	/	283
Title: AI learning	After 2025	131
Title: ai teacher education	From 2024 -2025	46
Title: ELIZA	From 1963 - 1973	2
Title: LLM as tutor	From 2024 - 2025	16
Title: large language learning	From 2024 - 2025	96

Tabelle 31 - ACM Suchbegriffe (PRISMA)

6.3.2 ArXiv

Suchbegriff	Filter	Anzahl
Title: AI education	/	304
Title: Education Inequities	/	13
Title: Human-Robot Interaction	/	531
Title: Humanoid robots ai	/	3
Title: NAO	/	36
Title: Anthropomorphism robots	/	27
Title: RAG Education	/	5
Title:Retrieval-Augmented Generation	/	989
Title: ranking BERT	/	21
Title: LLM learning	/	772
Title: Chunk Size	/	2
Title: Prompt Engineering	/	242
Title: effect learning robot	/	18
Title: chatbot prompt	/	12
Title: Foundation Models	/	2671
Title: hallucination llms	/	287
Title: GPT-4	/	251
Title: Text embeddings	/	440
Title: Approximate Nearest Neighbor	/	124
Title: RAG	/	693
Title: Humanoid robot	Past 12 Month	108
Title: Embedding models	/	1357

Tabelle 32 - ArXiv Suchbegriffe (PRISMA)

6.3.3 IEEE

Suchbegriff	Filter	Anzahl
Title: Artificial Intelligence education	/	459
Title: learning rag	/	14
Title: Chatbot higher education	/	16
Title: AI Learning Adaptive	/	53
Title: NAO	/	235
Title: Human Robot Interaction survey	2008-2008	12

Tabelle 33 - IEEE Suchbegriffe (PRISMA)

6.3.4 ScienceDirect

Suchbegriff	Filter	Anzahl
Title, abstract, keywords: education inequality africa	/	296
Title: Artificial intelligence education	/	326
Title: chatbots education	/	66
Title: uncanny valley	/	47
Title: Retrieval-augmented generation	/	72
Title: generative AI	/	710
Title: chatbot empathy	/	11
Title: human-robot interaction	/	410
Title: Technology acceptance model	/	374
Title: theory of planned behavior	Ajzen	5

Title: natural language processing	2024, 2025	488
Title: chatbots	2023, 2024, 2025	793
Title: UTAUT education		8

Tabelle 34 - ScienceDirect Suchbegriffe (PRISMA)

6.3.5 Springer

Suchbegriff	**Filter**	**Anzahl**
Title: ungleichheiten lehrkräfte	/	243
Title: Technology education	From 2023 - 2025	671
Title: Humanoid robot learning	Last 12 Months	821
Title: AI in education chatbots	Last 12 Months	12
Title: Artificial Intelligence in Education	Last 24 Months	378
Title: Humanoid robots school	/	2
Title: Learning Artificial Intelligence	/	336
Title: AI chatbots	Last 12 Months	123
Title: educational chatbot	/	20
Title: humanoid robots education	/	6
Title: NAO	From 2021 - 2025	78
Title: Uncanny Valley	/	68
Title: chatbot overview	/	1
Title: Retrieval-Augmented Generation	/	124
Title: effect chatbot	/	16
Title: future chatbot	/	6
Title: Anthropomorphism robots	/	21
Title: hallucination large language models	/	13

Title: large language models learning	/	131
Title: Large Language Models education	/	60
Title: LLM education	/	27
Title: Human-Robot interaction emotion	/	17
Title: Soziale Robotik	/	8
Title: chatbots education	/	65
Title: robots sensor	/	41
Title: llm learning	/	52
Title: university chatbots	2024-2025	6

Tabelle 35 - Springer Suchbegriffe (PRISMA)

6.3.6 Google Scholar

Suchbegriff	Filter	Anzahl
allintitle: Belief, attitude, intention and behaviour: An introduction to theory and research	/	1
allintitle: embodiment social presence	/	16
allintitle: NAO university	/	16
allintitle: social robot children	/	264
allintitle: SUS	author:brooke	5
allintitle: large language model teaching acceptance	/	1
allintitle: User acceptance of information technology Venkatesh	/	1
allintitle: Social Presence robots	/	31
alintitle: learning chatbots	/	78
allintitle: Technology acceptance model author:Davis	/	8

Tabelle 36 - GoogleScholar Suchbegriffe (PRISMA)

6.3.7 Taylor & Francis

Suchbegriff	Filter	Anzahl
Title: teacher shortage	01/01/2022 TO 12/31/2025	62
Title: humanoid robot education	/	3

Tabelle 37 - Taylor&Francis Suchbegriffe (PRISMA)

6.3.8 SpringerOpen

Suchbegriff	Filter	Anzahl
AI K-12	/	112
humanoid robots	/	116
STEM education robotics	/	175
chatbots in education	/	218

Tabelle 38 - SpringerOpen Suchbegriffe (PRISMA)

6.3.9 Nature

Suchbegriff	Filter	Anzahl
Title: Uncanny Valley	/	42
Title:LLMs	/	150

Tabelle 39 - Nature Suchbegriffe (PRISMA)

6.3.10 MDPI

Suchbegriff	Filter	Anzahl
Learning Artificial Intelligence	Journal: Education Sciences	3
Retrieval Augmented Generation education	/	1

Tabelle 40 - MDPI Suchbegriffe (PRISMA)

6.3.11 Frontsier

Suchbegriff	Filter	Anzahl
uncanny valley	Top 3 articles	3

Tabelle 41 - Frontsier Suchbegriffe (PRISMA)

6.3.12 Google Suche

Suchbegriff	Filter	Anzahl
Agenda 2030	Region: austria	128
LlamaIndex	First page	10
Streamlit	First page	10
NAOqi softbank	First page	10
RASA	First page	10
nlmatics	First page	10
fastapi	First page	10

Tabelle 42 - Agenda Suchbegriffe (PRISMA)

6.3.13 Externe Quellen

In der vorliegenden Arbeit wurden insgesamt 187 Quellen berücksichtigt, die durch eine systematische Recherche in einschlägigen wissenschaftlichen Publikationen identifiziert wurden. Die Auswahl der Literatur erfolgte dabei überwiegend über Referenzlisten und Zitationsangaben anderer relevanter Fachartikel. Auf diese Weise konnte ein breites Spektrum an einschlägigen Studien erschlossen werden, dass die thematische Grundlage für die Analyse und Diskussion dieser Arbeit bildet.

6.4 PRISMA 2020 Flow Chart

Die folgende PRISMA-2020-Flowchart veranschaulicht die einzelnen Komponenten.

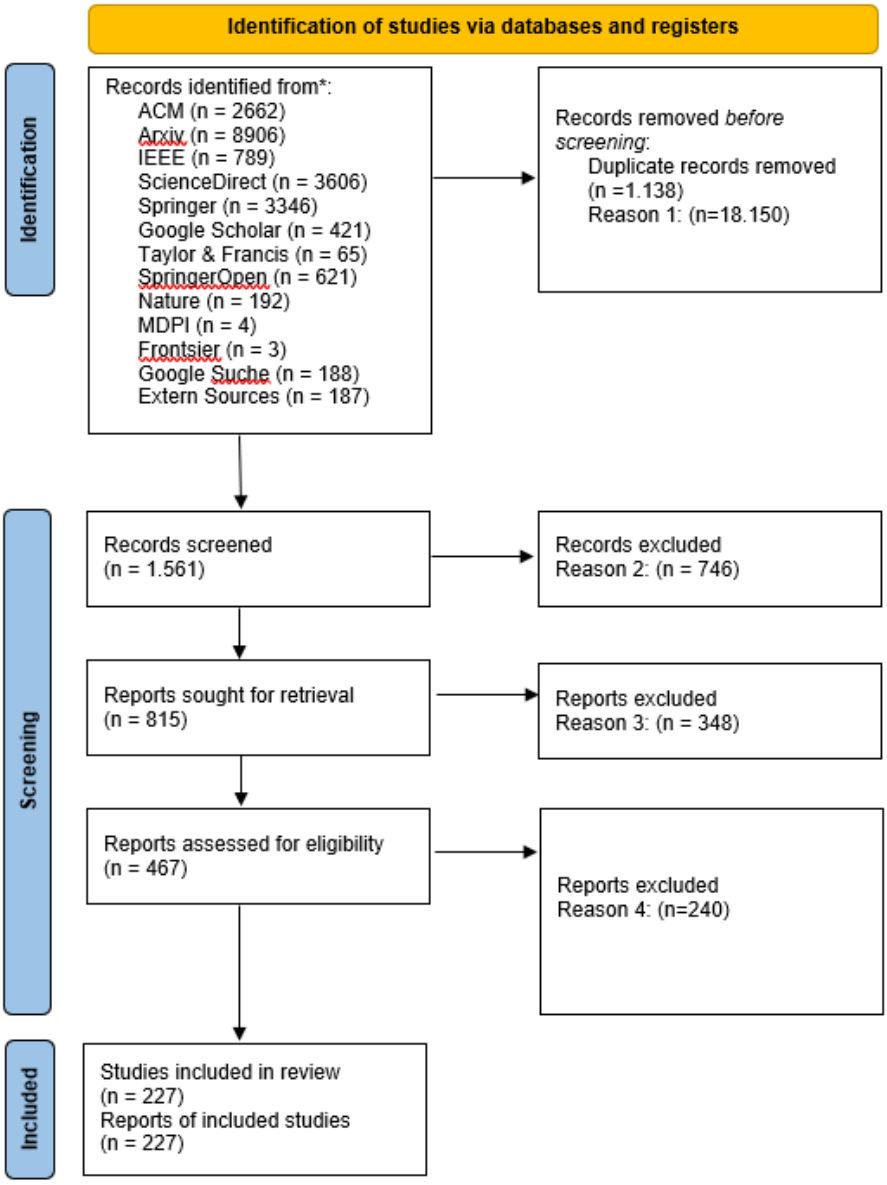

Abbildung 30 - PRISMA 2020 Diagramm

6.4.1 Identifikation

Die Identifikation relevanter Literatur erfolgte durch eine systematische Suche in anerkannten wissenschaftlichen Datenbanken. Dabei wurden gezielt Suchbegriffe (wie im Kapitel 6.3 beschrieben) verwendet, die das zentrale Forschungsthema der Diplomarbeit widerspiegeln.

6.4.2 Einschlusskriterium

Zur Sicherstellung einer hohen wissenschaftlichen Qualität wurden ausschließlich Artikel berücksichtigt, die:

- einen DOI-Schlüssel besitzen (außer Google Suche),
- in renommierten Verlagen veröffentlicht wurden,
- in begutachteten Fachzeitschriften erschienen sind.

6.4.3 Duplikate

Durch den Einsatz von Mendely und Zotero wurden 1.138 Duplikate mit dem gleichen Titel und DOI entdeckt.

Insgesamt wurden 18.150 Einträge ausgeschlossen, da deren Titel nicht mit der Forschungsfrage in Zusammenhang standen. Teilweise waren die verwendeten Suchbegriffe bewusst allgemein gewählt, um ein möglichst breites Spektrum potenziell relevanter Literatur zu erfassen.

6.4.4 Screening

Es verblieben 1561 Einträge, deren Abstracts oder Introduction geprüft wurden. Davon erwiesen sich 746 als nicht relevant und wurden vom weiteren Screening ausgeschlossen.

6.4.5 Retrieval

In diesem Prozess wurden 815 Studien einbezogen, 348 davon passten jedoch nach kurzer Durchsicht des Textes nicht zur Diplomarbeit.

6.4.6 Reports assessed for eligibility

In diesem Prozess blieben noch 467 wurde der Text genau und gründlich gelesen 240 davon gingen jedoch nicht genauer in die Tiefe.

6.4.7 Final Studies

Für diese Masterarbeit wurden 227 Quellen herangezogen. Angesichts der Breite des Forschungsfeldes wäre jedoch die Einbeziehung weiterer Literatur grundsätzlich möglich.

6.4.8 Ausschlusskriterium

Die Studie wurde aus folgenden Gründen ausgeschlossen:

Grund 1: Der Titel entspricht nicht der Fragestellung.

Grund 2: Das Abstract oder die Einleitung entspricht nicht den Kontext der Diplomarbeit.

Grund 3: Nach kurzer Sichtung des Volltextes erfüllt die Studie nicht die erforderlichen Kriterien.

Grund 4: Nach genauer Durchsicht des Volltextes erweist sich die Studie als nicht relevant für die Forschungsfrage.

7 Literaturverzeichnis

[1] A. Lieb and T. Goel, “Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education,” *Conference on Human Factors in Computing Systems - Proceedings*, May 2024, doi: 10.1145/3613905.3647957.

[2] K. Peyton, S. Unnikrishnan, and B. Mulligan, “A review of university chatbots for student support: FAQs and beyond,” *Discov. Educ.*, vol. 4, no. 1, Jan. 2025.

[3] J. Weizenbaum, “ELIZA—a computer program for the study of natural language communication between man and machine,” *Commun. ACM*, vol. 9, no. 1, pp. 36–45, Jan. 1966.

[4] W. Y. Leong, Y. Z. Leong, and W. S. Leong, “Artificial Intelligence in education,” in *IET International Conference on Engineering Technologies and Applications (ICETA 2024)*, 2024, pp. 183–184. doi: 10.1049/icp.2024.4341.

[5] H. Chen and C. C. Anyanwu, “AI in education: Evaluating the impact of moodle AI-powered chatbots and metacognitive teaching approaches on academic performance of higher Institution Business Education students,” *Educ. Inf. Technol.*, vol. 30, no. 9, pp. 12197–12212, Jun. 2025.

[6] D. Richter, Y. Huang, and E. Richter, “Ungleichheiten in der Lehrkräfteversorgung: Eine Analyse zur Verteilung qualifizierten Lehrpersonals auf Schulen mit unterschiedlicher Schülerschaft und verschiedenen sozio-ökonomischen Kontexten,” *Zeitschrift für Erziehungswissenschaft*, vol. 27, no. 6, pp. 1491–1517, 2024, doi: 10.1007/s11618-024-01271-2.

[7] A. Frola, M. Delprato, and A. Chudgar, “Lack of educational access, women’s empowerment and spatial education inequality for the Eastern and Western Africa regions,” *Int J Educ Dev*, vol. 104, p. 102939, 2024, doi: https://doi.org/10.1016/j.ijedudev.2023.102939.

[8] C. I. Pappa, D. Georgiou, and D. Pittich, “Technology education in primary schools: addressing teachers’ perceptions, perceived barriers, and needs,” *Int J Technol Des Educ*, vol. 34, no. 2, pp. 485–503, 2024, doi: 10.1007/s10798-023-09828-8.

[9] J. Jiang and S. Y. Yip, “Teacher shortage: an analysis of the rural teachers living subsidy policy on teacher attraction and retention in rural Western China,” *Asia-Pacific Journal of Teacher Education*, vol. 52, no. 3, pp. 316–331, May 2024, doi: 10.1080/1359866X.2024.2328682.

[10] P. Pholphirul, P. Rukumnuaykit, and S. Teimtad, “Teacher shortages and educational outcomes in developing countries: Empirical evidence from PISA-Thailand,” *Cogent Education*, vol. 10, no. 2, p. 2243126, Dec. 2023, doi: 10.1080/2331186X.2023.2243126.

[11] J. Hlongwane, G. N. Shava, A. Mangena, and T. Muzari, “Towards the integration of Artificial Intelligence in higher education, challenges and opportunities: The African context, a case of Zimbabwe,” *International Journal of Research and Innovation in Social Science*, vol. VIII, no. IIIS, pp. 417–435, 2024.

[12] A. N. T. Dieu, H. T. Nguyen, and C. T. D. Cong, “The enhanced context for AI-generated learning advisors with Advanced RAG,” in *2024 18th International Conference on Advanced Computing and Analytics (ACOMPA)*, Nov. 2024, pp. 94–101. doi: 10.1109/ACOMPA64883.2024.00021.

[13] S. Neupane, E. Hossain, J. Keith, H. Tripathi, F. Ghiasi, N. A. Golilarz, A. Amirlatifi, S. Mittal, and S. Rahimi, “From Questions to Insightful Answers: Building an Informed Chatbot for University Resources,” *arXiv preprint arXiv:2405.08120*, 2024.
[14] K. Y. Fung, L. H. Lee, K. F. Sin, S. Song, and H. Qu, “Humanoid robot-empowered language learning based on self-determination theory,” *Educ. Inf. Technol.*, vol. 29, no. 14, pp. 18927–18957, Oct. 2024.
[15] Z. Zhao, Z. Yin, J. Sun, and P. Hui, “Embodied AI-Guided Interactive Digital Teachers for Education,” *Proceedings - SIGGRAPH Asia 2024 Educator’s Forum, SA 2024*, Dec. 2024, doi: 10.1145/3680533.3697070.
[16] E. R. Ogbo-Gebhard and O. Ogbo, “Ogbo-Gebhardt, Erezi; Ogbo, Oruaro Standard-Nutzungsbedingungen: Using a Large Language Model-Powered Assistant in Teaching: Stories of Acceptance, Use, and Impact among Ethnic Minority Students,” 2024. [Online]. Available: https://hdl.handle.net/10419/302517
[17] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, “An LLM-Driven Chatbot in Higher Education for Databases and Information Systems,” *IEEE Transactions on Education*, pp. 1–14, 2024, doi: 10.1109/TE.2024.3467912.
[18] Venkatesh, Morris, Davis, and Davis, “User acceptance of information technology: Toward a unified view,” *MIS Q*, vol. 27, no. 3, p. 425, 2003.
[19] J. Brooke, “SUS – a quick and dirty usability scale,” 1996, pp. 189–194.
[20] D. Erol Barkana, M. Wahde, and M. Suvanto, “Interactive problem-solving with humanoid robots and non-expert users,” in *Proceedings of the 17th International Conference on Agents and Artificial Intelligence*, SCITEPRESS - Science and Technology Publications, 2025, pp. 870–879.
[21] A. Wedenborn, P. Wik, O. Engwall, and J. Beskow, “The effect of a physical robot on vocabulary learning,” 2019.
[22] A. Łukasik and A. Gut, “From robots to chatbots: unveiling the dynamics of human-AI interaction,” *Front. Psychol.*, vol. 16, p. 1569277, Apr. 2025.
[23] Q. Zhou, B. Li, L. Han, and M. Jou, “Talking to a bot or a wall? How chatbots vs. human agents affect anticipated communication quality,” *Comput Human Behav*, vol. 143, p. 107674, 2023, doi: https://doi.org/10.1016/j.chb.2023.107674.
[24] B. Tarlan and N. Erdal, “How can I assist you today?: A comparative analysis of a humanoid robot and a virtual human avatar in human perception,” 2024.
[25] “Agenda 2030.”
[26] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, and Z. Du, “Artificial intelligence in education: A systematic literature review,” *Expert Syst. Appl.*, vol. 252, no. 124167, p. 124167, Oct. 2024.
[27] X. Hu, S. Xu, R. Tong, and A. Graesser, “Generative AI in education: From foundational insights to the Socratic Playground for learning,” 2025.
[28] R. Alfredo, V. Echeverria, Y. Jin, L. Yan, Z. Swiecki, D. Gašević, and R. Martinez-Maldonado, “Human-centred Learning Analytics and AI in Education: A systematic literature review,” *arXiv [cs.CY]*, 2023.
[29] F. Kamalov, D. S. Calonge, D. Smail Linda and Azizov, D. R. Thadani, and A. Kwong Theresa and Atif, “Evolution of AI in Education: Agentic Workflows,” 2025.
[30] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, and G. Siemens, “A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour,” *Int. J. Educ. Technol. High. Educ.*, vol. 21, no. 1, Jan. 2024.

[31] W. Xu and F. Ouyang, “The application of AI technologies in STEM education: a systematic review from 2011 to 2021,” *Int J STEM Educ*, vol. 9, no. 1, p. 59, 2022, doi: 10.1186/s40594-022-00377-5.
[32] F. Ouyang and P. Jiao, “Artificial intelligence in education: The three paradigms,” *Computers and Education: Artificial Intelligence*, vol. 2, p. 100020, 2021, doi: https://doi.org/10.1016/j.caeai.2021.100020.
[33] L. Casal-Otero, A. Catala, C. Fernández-Morante, M. Taboada, B. Cebreiro, and S. Barro, “AI literacy in K-12: a systematic literature review,” *Int J STEM Educ*, vol. 10, no. 1, p. 29, 2023, doi: 10.1186/s40594-023-00418-7.
[34] “die hochschullehre,” 2024, *wbv Publikation*.
[35] R. Panda, “Artificial Intelligence in Educational Systems: From Early Computational Tools to Contemporary AI-Enhanced Learning Environments,” *International Journal for Research Publication and Seminars*, vol. 5, pp. 3756–3760, Aug. 2024, doi: 10.22271/23947519.2024.v10.i4a.2415https://doi.org/10.55248/gengpi.5.0824.2213.
[36] K. Holstein and S. Doroudi, “Equity and artificial intelligence in education: Will ``AIEd'' amplify or alleviate inequities in education?,” 2021.
[37] G. Fan, D. Liu, R. Zhang, and L. Pan, “The impact of AI-assisted pair programming on student motivation, programming anxiety, collaborative learning, and programming performance: a comparative study with traditional pair programming and individual approaches,” *Int J STEM Educ*, vol. 12, no. 1, p. 16, 2025, doi: 10.1186/s40594-025-00537-3.
[38] W. Cui, Z. Xue, and K.-P. Thai, “Performance comparison of an AI-based Adaptive Learning System in China,” 2019.
[39] D. Lee, Y. Huh, C.-Y. Lin, C. M. Reigeluth, and E. Lee, “Differences in personalized learning practice and technology use in high- and low-performing learner-centered schools in the United States,” *Educ. Technol. Res. Dev.*, vol. 69, no. 2, pp. 1221–1245, Feb. 2021.
[40] J. Henze, A. Bresges, and S. Becker-Genschow, “AI-supported data analysis boosts student motivation and reduces stress in physics education,” 2024.
[41] T. Netland, O. von Dzengelevski, K. Tesch, and D. Kwasnitschka, “Comparing human-made and AI-generated teaching videos: An experimental study on learning effects,” 2025.
[42] A. Bhutoria, “Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model,” *Computers and Education: Artificial Intelligence*, vol. 3, p. 100068, 2022, doi: https://doi.org/10.1016/j.caeai.2022.100068.
[43] D. Ifenthaler, R. Majumdar, P. Gorissen, M. Judge, S. Mishra, J. Raffaghelli, and A. Shimada, “Artificial intelligence in education: Implications for policymakers, researchers, and practitioners,” *Technol. Knowl. Learn.*, Jun. 2024.
[44] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education – where are the educators?,” *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.
[45] I. Gligorea, M. Cioca, R. Oancea, A.-T. Gorski, H. Gorski, and P. Tudorache, “Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review,” *Educ Sci (Basel)*, vol. 13, no. 12, 2023, doi: 10.3390/educsci13121216.
[46] N. Annamalai, R. A. Rashid, U. Munir Hashmi, M. Mohamed, and A. Harb Alqaryouti Marwan and Eddin Sadeq, “Using chatbots for English language learning in higher

education," *Computers and Education: Artificial Intelligence*, vol. 5, no. 100153, p. 100153, 2023.
[47] C. Chalmers, T. Keane, M. Boden, and M. Williams, "Humanoid robots go to school," *Educ. Inf. Technol.*, vol. 27, no. 6, pp. 7563–7581, Jul. 2022.
[48] I. Buchem and N. Bäcker, "HUMANOID ROBOTS IN HIGHER EDUCATION: AN EXPLORATORY STUDY ON APPLYING THE NAO ROBOT AS AN EDUCATIONAL TECHNOLOGY IN BUSINESS STUDIES IN HYBRID SETTINGS," Aug. 2022, pp. 8707–8714. doi: 10.21125/inted.2022.2264.
[49] X. Kong, H. Fang, W. Chen, J. Xiao, and M. Zhang, "Examining human–AI collaboration in hybrid intelligence learning environments: insight from the Synergy Degree Model," *Humanit Soc Sci Commun*, vol. 12, no. 1, p. 821, 2025, doi: 10.1057/s41599-025-05097-z.
[50] Y. Tu, J. Chen, and C. Huang, "Empowering personalized learning with generative artificial intelligence: Mechanisms, challenges and pathways," *Front. Digit. Educ.*, vol. 2, no. 2, Jun. 2025.
[51] P. Chittò, M. Baez, F. Daniel, and B. Benatallah, "Automatic generation of chatbots for conversational web browsing," in *Conceptual Modeling*, in Lecture notes in computer science. , Cham: Springer International Publishing, 2020, pp. 239–249.
[52] J. Sidlauskiene, Y. Joye, and V. Auruskeviciene, "AI-based chatbots in conversational commerce and their effects on product and price perceptions," *Electron. Mark.*, vol. 33, no. 1, p. 24, May 2023.
[53] L. Adamopoulou Eleni and Moussiades, "An Overview of Chatbot Technology," in *Artificial Intelligence Applications and Innovations*, L. and P. E. Maglogiannis Ilias and Iliadis, Ed., Cham: Springer International Publishing, 2020, pp. 373–383.
[54] Z. Cai, S. Park, N. Nixon, and S. Doroudi, "Advancing knowledge together: Integrating large language model-based conversational AI in small group collaborative learning," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, May 2024, pp. 1–9.
[55] M. Kurni, M. S. Mohammed, and K. G. Srinivasa, "Chatbots for education," in *A Beginner's Guide to Introduce Artificial Intelligence in Teaching and Learning*, Cham: Springer International Publishing, 2023, pp. 173–198.
[56] T. Ait Baha, M. El Hajji, Y. Es-Saady, and H. Fadili, "The impact of educational chatbot on student learning experience," *Educ. Inf. Technol.*, vol. 29, no. 8, pp. 10153–10176, Jun. 2024.
[57] M. A. Kuhail, N. Alturki, and K. Alramlawi Salwa and Alhejori, "Interacting with educational chatbots: A systematic review," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 973–1018, Jan. 2023.
[58] Y. Xu and M. Warschauer, "Young children's reading and learning with conversational agents," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, May 2019.
[59] S. Gupta and Y. Chen, "Supporting inclusive learning using chatbots? A chatbot-led interview study," *Journal of Information Systems Education*, vol. 33, no. 1, pp. 98–108, 2022.
[60] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100417, 2025, doi: https://doi.org/10.1016/j.caeai.2025.100417.

[61] L. Benotti, M. C. Mart\'\inez, and F. Schapachnik, “Engaging high school students using chatbots,” in *Proceedings of the 2014 conference on Innovation & technology in computer science education - ITiCSE ’14*, New York, New York, USA: ACM Press, 2014.

[62] N. Phaokla and P. Netinant, “Design an environment information chatbots system for a smart school framework,” in *2021 The 4th International Conference on Software Engineering and Information Management*, New York, NY, USA: ACM, Jan. 2021.

[63] J. J. Merelo, P. A. Castillo, A. M. Mora, F. Barranco, N. Abbas, and O. Guillén Alberto and Tsivitanidou, “Exploring the role of chatbots and messaging applications in higher education: A teacher’s perspective,” in *Learning and Collaboration Technologies. Novel Technological Environments*, in Lecture notes in computer science. , Cham: Springer International Publishing, 2022, pp. 205–223.

[64] Z. W. Taylor and S. Owusu, “Chatbot integration within United States higher education websites: Historical trends from 2017–2023,” *Technol. Knowl. Learn.*, Oct. 2024.

[65] M. Tamascelli, O. Bunch, B. Fowler, M. Taeb, and A. Cohen, “Academic advising chatbot powered with AI agent,” in *Proceedings of the 2025 ACM Southeast Conference*, New York, NY, USA: ACM, Apr. 2025, pp. 195–202.

[66] S. Jusoh, H. Al Fawareh, and H. Abdul Kadir Rabiah and Hosseinzadeh, “HelpBot: A web-based chatbot to handle depression among adolescents,” in *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, New York, NY, USA: ACM, Apr. 2024, pp. 149–154.

[67] M. Klesel and H. F. Wittmann, “Retrieval-augmented generation (RAG),” *Bus. Inf. Syst. Eng.*, Jun. 2025.

[68] L. Cao, “Humanoid robots and humanoid AI: Review, perspectives and directions,” 2024.

[69] S. Cavicchi, A. Abubshait, G. Siri, M. Mustile, and F. Ciardo, “Can humanoid robots be used as a cognitive offloading tool?,” *Cogn. Res. Princ. Implic.*, vol. 10, no. 1, p. 17, Apr. 2025.

[70] Y.-C. Chen, S.-L. Yeh, W. Lin, H.-P. Yueh, and L.-C. Fu, “The effects of social presence and familiarity on children-robot interactions,” *Sensors (Basel)*, vol. 23, no. 9, Apr. 2023.

[71] A. Asghar, A. Patra, and K. S. Ravi, “The potential scope of a humanoid robot in anatomy education: a review of a unique proposal,” *Surg. Radiol. Anat.*, vol. 44, no. 10, pp. 1309–1317, Oct. 2022.

[72] D. Fu, F. Abawi, P. Allgeuer, and S. Wermter, “Human impression of humanoid robots mirroring social cues,” in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA: ACM, Mar. 2024, pp. 458–462.

[73] C.-C. Ho and K. F. MacDorman, “Measuring the uncanny valley effect,” *Int. J. Soc. Robot.*, vol. 9, no. 1, pp. 129–139, Jan. 2017.

[74] A. Diel, S. Weigelt, and K. F. Macdorman, “A meta-analysis of the uncanny valley’s independent and dependent variables,” *ACM Trans. Hum. Robot Interact.*, vol. 11, no. 1, pp. 1–33, Mar. 2022.

[75] B. Kim, E. de Visser, and E. Phillips, “Two uncanny valleys: Re-evaluating the uncanny valley across the full spectrum of real-world human-like robots,” *Comput Human Behav*, vol. 135, p. 107340, 2022, doi: https://doi.org/10.1016/j.chb.2022.107340.

[76] R. K. Moore, “A Bayesian explanation of the ‘Uncanny Valley’ effect and related psychological phenomena,” *Sci Rep*, vol. 2, no. 1, p. 864, 2012, doi: 10.1038/srep00864.
[77] J. Kätsyri, K. Förger, M. Mäkäräinen, and T. Takala, “A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness,” *Front. Psychol.*, vol. 6, p. 390, Apr. 2015.
[78] D. Gouaillier, V. Hugel, C. Blazevic Pierre and Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, “Mechatronic design of NAO humanoid,” in *2009 IEEE International Conference on Robotics and Automation*, IEEE, May 2009.
[79] A. Paraschos, N. I. Spanoudakis, and M. G. Lagoudakis, “Model-driven behavior specification for robotic teams,” in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, in AAMAS ’12. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 171–178.
[80] H. Fadli, E. Hidayat, and C. Machbub, “Design and implementation of walking pattern and trajectory compensator of NAO humanoid robot,” in *2016 6th International Conference on System Engineering and Technology (ICSET)*, 2016, pp. 184–188. doi: 10.1109/ICSEngT.2016.7849647.
[81] S. R. Cruz-Ram\’\irez, M. Garc\’\ia-Mart\’\inez, and J. M. Olais-Govea, “NAO robots as context to teach numerical methods,” *Int. J. Interact. Des. Manuf. (IJIDeM)*, vol. 16, no. 4, pp. 1337–1356, Oct. 2022.
[82] N. Pöhner and M. Hennecke, “Evaluation of a robotics course with the humanoid robot NAO in CS teacher education,” in *Proceedings of the 13th Workshop in Primary and Secondary Computing Education*, in WiPSCE ’18. New York, NY, USA: Association for Computing Machinery, 2018. doi: 10.1145/3265757.3265786.
[83] P. B. Lyk and M. Lyk, “Nao as an Authority in the Classroom: Can Nao Help the Teacher to Keep an Acceptable Noise Level?,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, in HRI’15 Extended Abstracts. New York, NY, USA: Association for Computing Machinery, 2015, pp. 77–78. doi: 10.1145/2701973.2702014.
[84] H. Banaeian and I. Gilanlioglu, “Influence of the NAO robot as a teaching assistant on university students’ vocabulary learning and attitudes,” *Australasian Journal of Educational Technology*, pp. 71–87, Aug. 2021, doi: 10.14742/ajet.6130.
[85] L. Perla, L. S. Agrati, and I. Amanti, “Child-NAO robot interaction: A study from the educator mediation perspective,” in *Communications in Computer and Information Science*, in Communications in computer and information science. , Cham: Springer Nature Switzerland, 2025, pp. 340–348.
[86] Q. Yang, H. Lu, D. Liang, and H. Gong Shengrong and Feng, “Surprising performances of students with autism in classroom with NAO robot,” 2024.
[87] I. Buchem and N. Baecker, “NAO Robot as Scrum Master: Results from a scenario-based study on building rapport with a humanoid robot in hybrid higher education settings,” in *Training, Education, and Learning Sciences*, AHFE International, 2022.
[88] A. Pande and D. Mishra, “Humanoid robot as an educational assistant – insights of speech recognition for online and offline mode of teaching,” *Behaviour & Information Technology*, vol. 44, no. 5, pp. 975–992, 2025, doi: 10.1080/0144929X.2024.2344726.
[89] F. Ouyang and W. Xu, “The effects of educational robotics in STEM education: a multilevel meta-analysis,” *Int J STEM Educ*, vol. 11, no. 1, p. 7, 2024, doi: 10.1186/s40594-024-00469-4.

[90] S. Ekström and L. Pareto, “The dual role of humanoid robots in education: As didactic tools and social actors,” *Educ. Inf. Technol.*, vol. 27, no. 9, pp. 12609–12644, Nov. 2022.
[91] H. S. Yun, V. Taliaronak, M. Kirtay, J. Chevelère, H. Hübert, V. V Hafner, N. Pinkwart, and R. Lazarides, “Challenges in designing teacher robots with motivation based gestures,” 2023.
[92] H. Ates and M. Polat, “Exploring adoption of humanoid robots in education: UTAUT-2 and TOE models for science teachers,” *Educ. Inf. Technol.*, vol. 30, no. 9, pp. 12765–12806, Jun. 2025.
[93] Y. Liu, X. Hu, S. Zhang, J. Chen, F. Wu, and F. Wu, “Fine-grained Guidance for retrievers: Leveraging LLMs’ feedback in retrieval-Augmented Generation,” 2024.
[94] V. Klotzman, C. V Lopes, J. Schomberg, Y. S. Armanyous, D. Linden, I. Ma, A. Giron, P. Yu, H. Ahmad, L. F. Goodman, M. Kabeer, and Y. Guner, “Development of a CDH-specific conversational assistant using RAG,” in *Communications in Computer and Information Science*, in Communications in computer and information science. , Cham: Springer Nature Switzerland, 2025, pp. 31–46.
[95] C. Sharma, “Retrieval-Augmented Generation: A comprehensive survey of architectures, enhancements, and robustness frontiers,” 2025.
[96] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, “Advancements in natural language processing: Implications, challenges, and future directions,” *Telematics and Informatics Reports*, vol. 16, p. 100173, 2024, doi: https://doi.org/10.1016/j.teler.2024.100173.
[97] L. I. D. Faruk, S. Funilkul, P. Mongkolnam, P. Puengwattanapong, and D. Pal, “Exploring User Experience with Voice Assistants: Impact of Prior Experience on Voice Assistants,” in *Proceedings of the 13th International Conference on Advances in Information Technology*, in IAIT ’23. New York, NY, USA: Association for Computing Machinery, 2023. doi: 10.1145/3628454.3629470.
[98] S. Pandey and S. Sharma, “A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning,” *Healthcare Analytics*, vol. 3, p. 100198, 2023, doi: https://doi.org/10.1016/j.health.2023.100198.
[99] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, “Developing Retrieval Augmented Generation (RAG) based LLM systems from PDFs: An experience report,” 2024.
[100] Y. Gao, Y. Xiong, X. Gao, J. Jia Kangxiang and Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-Augmented Generation for large Language Models: A survey,” 2023.
[101] “Advancing retrieval-augmented generation for enhanced QA performance: A multi-agent ChatPDF approach,” *International Research Journal of Modernization in Engineering Technology and Science*, Jun. 2025.
[102] A. Brown, M. Roman, and B. Devereux, “A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges,” 2025.
[103] S. Dakshit, “Faculty Perspectives on the Potential of RAG in Computer Science Higher Education,” *The 25th Annual Conference on Information Technology Education*, pp. 19–24, Oct. 2024, doi: 10.1145/3686852.3686864.
[104] J. Swacha and M. Gracel, “Retrieval-Augmented Generation (RAG) chatbots for education: A survey of applications,” *Appl. Sci. (Basel)*, vol. 15, no. 8, p. 4234, Apr. 2025.

[105] Y.-K. and T. Y.-C. and S. S. Ko Hsing-Tzu and Liu, “Enhancing Python Learning Through Retrieval-Augmented Generation: A Theoretical and Applied Innovation in Generative AI Education,” in *Innovative Technologies and Learning*, M. and B. E. and H. Y.-M. Cheng Yu-Ping and Pedaste, Ed., Cham: Springer Nature Switzerland, 2024, pp. 164–173.
[106] Z. Levonian, C. Li, W. Zhu, A. Gade, O. Henkel, M.-E. Postle, and W. Xing, “Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference,” 2023.
[107] Y. Chu, P. He, H. Li, H. Han, K. Yang, Y. Xue, T. Li, and J. Krajcik Joseph and Tang, “Enhancing LLM-based short answer grading with retrieval-augmented generation,” 2025.
[108] Y. Liu, Q. Yang, J. Tang, C. Guo Tiezheng and Wang, P. Li, S. Xu, X. Gao, Z. Li, J. Liu, and Y. Wen, “Reducing hallucinations of large language models via hierarchical semantic piece,” *Complex Intell. Syst.*, vol. 11, no. 5, May 2025.
[109] X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman, and Y. Su, “Applications of natural language processing and large language models in materials discovery,” *Npj Comput. Mater.*, vol. 11, no. 1, Mar. 2025.
[110] M. Ding, S. Dong, and R. Grewal, “Generative AI and usage in marketing classroom,” *Cust. Needs Solut.*, vol. 11, no. 1, Dec. 2024.
[111] R. Langenderfer, “Large Language Magic: Conjuring the Future of Education with LLMs,” in *Computational Science and Computational Intelligence*, L. and S. F. and A. S. and G. M. F. Arabnia Hamid R. and Deligiannidis, Ed., Cham: Springer Nature Switzerland, 2025, pp. 159–164.
[112] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, “The role of large language models in personalized learning: a systematic review of educational impact,” *Discov. Sustain.*, vol. 6, no. 1, Apr. 2025.
[113] W. Xing, N. Nixon, S. Crossley, P. Denny, A. Lan, J. Stamper, and Z. Yu, “The Use of Large Language Models in Education,” *Int J Artif Intell Educ*, vol. 35, no. 2, pp. 439–443, 2025, doi: 10.1007/s40593-025-00457-x.
[114] X. J. Wang, C. P. Lee, and B. Mutlu, “LearnMate: Enhancing Online Education with LLM-Powered Personalized Learning Plans and Support,” in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, in CHI EA ’25. New York, NY, USA: Association for Computing Machinery, 2025. doi: 10.1145/3706599.3719857.
[115] S. E. Huber, K. Kiili, S. Nebel, R. M. Ryan, M. Sailer, and M. Ninaus, “Leveraging the Potential of Large Language Models in Education Through Playful and Game-Based Learning,” *Educ Psychol Rev*, vol. 36, no. 1, p. 25, 2024, doi: 10.1007/s10648-024-09868-z.
[116] V. Rus and P. Kendeou, “Are LLMs actually good for learning?,” *AI Soc*, 2025, doi: 10.1007/s00146-025-02323-9.
[117] L. Masanneck, S. G. Meuth, and M. Pawlitzki, “Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology,” *NPJ Digit. Med.*, vol. 8, no. 1, p. 137, Mar. 2025.
[118] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, G. Herbert-Voss Ariel and Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” 2020.

[119] L. Ouyang, J. Wu, X. Jiang, C. L. Almeida Diogo and Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, J. Ray Alex and Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” 2022.

[120] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, X. Tang Jiakai and Chen, Y. Lin, W. X. Zhao, and J. Wei Zhewei and Wen, “A survey on large language model based autonomous agents,” *Front. Comput. Sci.*, vol. 18, no. 6, Dec. 2024.

[121] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, J. Hsu Kyle and Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna Ranjay and Kuditipudi, A. Kumar, M. Ladhak Faisal and Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, A. Ma Tengyu and Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, A. Nair Suraj and Narayan, D. Narayanan, A. Newman Ben and Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, E. Piech Chris and Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, M. Xie Sang Michael and Yasunaga, J. You, M. Zaharia Matei and Zhang, T. Zhang, *et al.*, “On the opportunities and risks of foundation models,” 2021.

[122] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of Hallucination in Natural Language Generation,” *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023, doi: 10.1145/3571730.

[123] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, P. Kamar Ece and Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” 2023.

[124] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[125] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners,” 2022.

[126] Q. Fang, D. Nguyen, and D. L. Oberski, “Evaluating the construct validity of text embeddings with application to survey questions,” *EPJ Data Sci*, vol. 11, no. 1, p. 39, 2022, doi: 10.1140/epjds/s13688-022-00353-7.

[127] J. Cao, J. Fang, Z. Meng, and S. Liang, “Knowledge Graph Embedding: A Survey from the Perspective of Representation Spaces,” *ACM Comput. Surv.*, vol. 56, no. 6, Mar. 2024, doi: 10.1145/3643806.

[128] J. Cao, J. Fang, Z. Meng, and S. Liang, “Knowledge graph embedding: A survey from the perspective of representation spaces,” 2022.

[129] P. Aceves and J. A. Evans, “Mobilizing conceptual spaces: How word embedding models can inform measurement and theory within organization science,” *Organ. Sci.*, vol. 35, no. 3, pp. 788–814, May 2024.

[130] K. Blagec, H. Xu, A. Agibetov, and M. Samwald, “Neural sentence embedding models for semantic similarity estimation in the biomedical domain,” *BMC Bioinformatics*, vol. 20, no. 1, p. 178, 2019, doi: 10.1186/s12859-019-2789-2.
[131] W. Wang, J. Wang, X. Peng, Y. Yang, C. Xiao, S. Yang, M. Wang, L. Wang, L. Li, and X. Chang, “Exploring best-matched embedding model and classifier for charging-pile fault diagnosis,” *Cybersecurity*, vol. 6, no. 1, p. 7, 2023, doi: 10.1186/s42400-023-00138-z.
[132] S. Kim, H. Song, H. Seo, and H. Kim, “Optimizing retrieval strategies for financial question answering documents in retrieval-Augmented Generation systems,” 2025.
[133] A. Xu, T. Yu, M. Du, P. Gundecha, Y. Guo, X. Zhu, M. Wang, P. Li, and X. Chen, “Generative AI and retrieval-augmented generation (RAG) systems for enterprise,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Oct. 2024, pp. 5599–5602.
[134] S. Prabhune and D. J. Berndt, “Deploying large language models with retrieval Augmented Generation,” 2024.
[135] P. Xu, W. Ping, X. Wu, C. McAfee Lawrence and Zhu, Z. Liu, S. Subramanian, E. Bakhturina, M. Shoeybi, and B. Catanzaro, “Retrieval meets long context large language models,” 2023.
[136] R. and J. L. and M. K. and M. S. and N.-O. I. and W. R. and L. A. and F. B. Wiratunga Nirmalie and Abeyratne, “CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering,” in *Case-Based Reasoning Research and Development*, M. G. and B. D. Recio-Garcia Juan A. and Orozco-del-Castillo, Ed., Cham: Springer Nature Switzerland, 2024, pp. 445–460.
[137] S. Ghanbari Haez, M. Segala, P. Bellan, S. Magnolini, L. Sanna, M. Consolandi, and M. Dragoni, “A Retrieval-Augmented Generation Strategy to Enhance Medical Chatbot Reliability,” in *Artificial Intelligence in Medicine: 22nd International Conference, AIME 2024, Salt Lake City, UT, USA, July 9–12, 2024, Proceedings, Part I*, Berlin, Heidelberg: Springer-Verlag, 2024, pp. 213–223. doi: 10.1007/978-3-031-66538-7_22.
[138] R. Nogueira and K. Cho, “Passage re-ranking with BERT,” 2019.
[139] J. Lin, R. Nogueira, and A. Yates, “Pretrained transformers for text ranking: BERT and beyond,” 2020.
[140] L. Xiong, C. Xiong, Y. Li, J. Tang Kwok-Fung and Liu, P. Bennett, J. Ahmed, and A. Overwijk, “Approximate Nearest Neighbor negative contrastive learning for dense text retrieval,” 2020.
[141] R. Upadhyay and M. Viviani, “Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy,” *Discov Computing*, vol. 28, no. 1, Apr. 2025.
[142] L. Masanneck, S. G. Meuth, and M. Pawlitzki, “Evaluating base and retrieval augmented LLMs with document or online support for evidence based neurology,” *NPJ Digit. Med.*, vol. 8, no. 1, p. 137, Mar. 2025.
[143] K. Ronanki, S. Arvidsson, and J. Axell, “Prompt engineering guidelines for using Large Language Models in Requirements Engineering,” 2025.
[144] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, H. Srivastava Ashay and Da Costa, S. Gupta, I. Rogers Megan L and Goncearenco, G. Sarli, D. Galynker Igor and Peskoff, M. Carpuat, J. White, S.

Anadkat, A. Hoyle, and P. Resnik, “The prompt report: A systematic survey of prompt engineering techniques,” 2024.
[145] C. Koyuturk, E. Theophilou, G. Patania Sabrina and Donabauer, A. Martinenghi, C. Antico, A. Telari, A. Testa, S. Bursic, F. Garzotto, D. Hernandez-Leo, U. Kruschwitz, D. Taibi, S. Amenta, M. Ruskov, and D. Ognibene, “Understanding learner-LLM chatbot interactions and the impact of prompting guidelines,” 2025.
[146] B. Yang, M. A. Al Mamun, J. M. Zhang, and G. Uddin, “Hallucination detection in large Language Models with metamorphic relations,” *Proc. ACM Softw. Eng.*, vol. 2, no. FSE, pp. 425–445, Jun. 2025.
[147] E. Oro, F. M. Granata, and M. Ruffolo, “A comprehensive evaluation of embedding models and LLMs for IR and QA across English and Italian,” *Big Data Cogn. Comput.*, vol. 9, no. 5, p. 141, May 2025.
[148] Z. Yu, S. Liu, P. Denny, A. Bergen, and M. Liut, “Integrating Small Language Models with Retrieval-Augmented Generation in Computing Education: Key Takeaways, Setup, and Practical Insights,” in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, in SIGCSETS 2025. New York, NY, USA: Association for Computing Machinery, 2025, pp. 1302–1308. doi: 10.1145/3641554.3701844.
[149] T. Kim, J. Springer, A. Raghunathan, and M. Sap, “Mitigating bias in RAG: Controlling the embedder,” 2025.
[150] S. R. Bhat, M. Rudat, J. Spiekermann, and N. Flores-Herr, “Rethinking chunk size for long-document retrieval: A multi-dataset analysis,” 2025.
[151] C. Merola and J. Singh, “Reconstructing context: Evaluating advanced chunking strategies for retrieval-augmented generation,” 2025.
[152] Z. Zhong, H. Liu, X. Cui, X. Zhang, and Z. Qin, “Mix-of-granularity: Optimize the chunking granularity for retrieval-Augmented Generation,” 2024.
[153] T. Koga, R. Wu, and K. Chaudhuri, “Privacy-preserving retrieval-augmented generation with differential privacy,” 2024.
[154] B. An, S. Zhang, and M. Dredze, “RAG LLMs are not safer: A safety analysis of Retrieval-Augmented Generation for large language models,” 2025.
[155] L. He, P. Tang, Y. Zhang, P. Zhou, and S. Su, “Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation,” *Inf Process Manag*, vol. 62, no. 4, p. 104150, 2025, doi: https://doi.org/10.1016/j.ipm.2025.104150.
[156] S. R. Kandula, “Securing retrieval-augmented generation - privacy risks and mitigation strategies,” *SSRN Electron. J.*, 2025.
[157] X. Xu, H. Weytjens, D. Zhang, I. Lu Qinghua and Weber, and L. Zhu, “RAGOps: Operating and managing retrieval-augmented generation pipelines,” 2025.
[158] S. Pahune, Z. Akhtar, V. Mandapati, and K. Siddique, “The importance of AI data governance in large language models,” Apr. 2025.
[159] A. Følstad, T. Araujo, P. B. Law Effie Lai-Chong and Brandtzaeg, S. Papadopoulos, L. Reis, M. Baez, G. Laban, P. McAllister, C. Ischen, R. Wald, F. Catania, R. Meyer von Wolff, S. Hobert, and E. Luger, “Future directions for chatbot research: an interdisciplinary research agenda,” *Computing*, vol. 103, no. 12, pp. 2915–2942, Dec. 2021.
[160] D. Wang Jiefei and Herath, “What Makes Robots? Sensors, Actuators, and Algorithms,” in *Foundations of Robotics: A Multidisciplinary Approach with Python and ROS*, D. Herath Damith and St-Onge, Ed., Singapore: Springer Nature Singapore, 2022, pp. 177–203. doi: 10.1007/978-981-19-1983-1_7.

[161] X. Zhong, H. Xin, W. Li, and M.-H. Zhan Zehui and Cheng, “The Design and application of RAG-based conversational agents for collaborative problem solving,” in *Proceedings of the 2024 9th International Conference on Distance Education and Learning*, New York, NY, USA: ACM, Jun. 2024, pp. 62–68.
[162] M. A. Goodrich and A. C. Schultz, *Human-Robot Interaction: A Survey*. Hanover, MA, USA: Now Publishers Inc., 2008.
[163] F. Wang Yue and Zhang, “Introduction,” in *Trends in Control and Decision-Making for Human–Robot Collaboration Systems*, F. Wang Yue and Zhang, Ed., Cham: Springer International Publishing, 2017, pp. 1–13. doi: 10.1007/978-3-319-40533-9_1.
[164] N. Mavridis, “A Review of Verbal and Non-Verbal Human-Robot Interactive Communication,” 2014.
[165] M. Farajtabar and M. Charbonneau, “The path towards contact-based physical human–robot interaction,” *Rob Auton Syst*, vol. 182, p. 104829, 2024, doi: https://doi.org/10.1016/j.robot.2024.104829.
[166] A.-N. Sharkawy, “Human-Robot Interaction: Applications,” *arXiv [cs.RO]*, 2021.
[167] I. H. Han, D. H. Kim, K. H. Nam, J. I. Lee, K.-H. Kim, J.-H. Park, and H. S. Ahn, “Human-robot interaction and social robot: The emerging field of healthcare robotics and current and future perspectives for spinal care,” *Neurospine*, vol. 21, no. 3, pp. 868–877, Sep. 2024.
[168] A. Billard and D. Grollman, “Human-robot interaction,” in *Encyclopedia of the Sciences of Learning*, Boston, MA: Springer US, 2012, pp. 1474–1476.
[169] Z. R. Khavas, R. Ahmadzadeh, and P. Robinette, “Modeling trust in human-robot interaction: A survey,” 2020.
[170] R. Stock-Homburg, “Survey of Emotions in Human–Robot Interactions: Perspectives from Robotic Psychology on 20 Years of Research,” *Int J Soc Robot*, vol. 14, no. 2, pp. 389–411, 2022, doi: 10.1007/s12369-021-00778-6.
[171] E. S. Cross, R. Hortensius, and A. Wykowska, “From social brains to social robots: applying neurocognitive insights to human-robot interaction,” *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 374, no. 1771, p. 20180024, Apr. 2019.
[172] H. Kamide, F. Eyssel, and T. Arai, “Psychological anthropomorphism of robots,” in *Social Robotics*, in Lecture notes in computer science. , Cham: Springer International Publishing, 2013, pp. 199–208.
[173] Y. Jung and S. Hahn, “Social robots as companions for lonely hearts: The role of anthropomorphism and robot appearance,” *arXiv [cs.RO]*, 2023.
[174] S. Thellman, M. de Graaf, and T. Ziemke, “Mental state attribution to robots: A systematic review of conceptions, methods, and findings,” *ACM Trans. Hum. Robot Interact.*, vol. 11, no. 4, pp. 1–51, Dec. 2022.
[175] M. Romeo, P. E. McKenna, D. A. Robb, G. Rajendran, B. Nesset, A. Cangelosi, and H. Hastie, “Exploring theory of mind for human-robot collaboration,” in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2022, pp. 461–468.
[176] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, “Theory of mind (ToM) on robots: a functional neuroimaging study,” in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, in HRI ’08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 335–342. doi: 10.1145/1349822.1349866.

[177] C. Gena, F. Manini, A. Lieto, A. Lillo, and F. Vernero, “Can empathy affect the attribution of mental states to robots?,” in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, New York, NY, USA: ACM, Oct. 2023, pp. 94–103.
[178] J. H. Chin, K. S. Haring, and P. Kim, “Understanding the neural mechanisms of empathy toward robots to shape future applications,” *Front. Neurorobot.*, vol. 17, p. 1145989, Apr. 2023.
[179] Z. R. Khavas, “A Review on Trust in Human-Robot Interaction,” 2021.
[180] D. Ullrich, A. Butz, and S. Diefenbach, “The development of overtrust: An empirical simulation and psychological analysis in the context of human-robot interaction,” *Front. Robot. AI*, vol. 8, p. 554578, Apr. 2021.
[181] T. Maeda and A. Quan-Haase, “When human-AI interactions become parasocial: Agency and anthropomorphism in affective design,” in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA: ACM, Jun. 2024.
[182] Y. Jung, M. Kwan, and K. Lee, “Effects of physical embodiment on social presence of social robots,” *Proceedings of Presence*, Aug. 2004.
[183] M. Mara and B. Leichtmann, “Soziale Robotik und Roboterpsychologie,” in *Soziale Roboter*, Wiesbaden: Springer Fachmedien Wiesbaden, 2021, pp. 169–189.
[184] R. J. Holden and B.-T. Karsh, “The technology acceptance model: its past and its future in health care,” *J. Biomed. Inform.*, vol. 43, no. 1, pp. 159–172, Feb. 2010.
[185] I. Ajzen, “The theory of planned behavior,” *Organ Behav Hum Decis Process*, vol. 50, no. 2, pp. 179–211, Dec. 1991, doi: 10.1016/0749-5978(91)90020-T.
[186] F. D. Davis and others, “Technology acceptance model: TAM,” *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption*, vol. 205, no. 219, p. 5, 1989.
[187] M. Fishbein and I. Ajzen, *Belief, attitude, intention and behaviour: An introduction to theory and research*, vol. 27. 1975.
[188] P. Legris, J. Ingham, and P. Collerette, “Why do people use information technology? A critical review of the technology acceptance model,” *Information & Management*, vol. 40, no. 3, pp. 191–204, 2003, doi: https://doi.org/10.1016/S0378-7206(01)00143-4.
[189] Venkatesh, Thong, and Xu, “Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology,” *MIS Q*, vol. 36, no. 1, p. 157, 2012.
[190] “DIN EN ISO 9241-11:2018-11, Ergonomie der Mensch-System-Interaktion_- Teil_11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO_9241-11:2018); Deutsche Fassung EN_ISO_9241-11:2018,” Berlin, 2019.
[191] J. Nielsen, *Usability Engineering*. in Interactive Technologies. Oxford, England: Morgan Kaufmann, 1994.
[192] B. Klug, “An overview of the system usability scale in library website and system usability testing,” *Weav. J. Libr. User Exp.*, vol. 1, no. 6, Apr. 2017.
[193] J. Brooke, “SUS: a retrospective,” *J. Usability Studies*, vol. 8, no. 2, pp. 29–40, Feb. 2013.
[194] A. Revythi and N. Tselios, “Extension of Technology Acceptance Model by using System Usability Scale to assess behavioral intention to use e-learning,” 2017.
[195] “SUS: A ‘quick and dirty’ usability scale,” in *Usability Evaluation In Industry*, CRC Press, 1996, pp. 207–212.
[196] T. Debets, S. K. Banihashem, D. Joosten-Ten Brinke, T. E. J. Vos, G. Maillette de Buy Wenniger, and G. Camp, “Chatbots in education: A systematic review of objectives,

underlying technology and theory, evaluation criteria, and impacts," *Comput Educ*, vol. 234, p. 105323, 2025, doi: https://doi.org/10.1016/j.compedu.2025.105323.
[197] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *Int. J. Educ. Technol. High. Educ.*, vol. 20, no. 1, Oct. 2023.
[198] M. Laun and F. Wolff, "Chatbots in education: Hype or help? A meta-analysis," *Learn Individ Differ*, vol. 119, p. 102646, 2025, doi: https://doi.org/10.1016/j.lindif.2025.102646.
[199] R. Guan, M. Raković, G. Chen, and D. Ga\v sević, "How educational chatbots support self-regulated learning? A systematic review of the literature," *Educ. Inf. Technol.*, vol. 30, no. 4, pp. 4493–4518, Mar. 2025.
[200] B. Mirzababaei, K. Maitz, A. Fessl, and V. Pammer-Schindler, "Interactive web-based learning materials vs. Tutorial chatbot: Differences in user experience," in *Lecture Notes in Computer Science*, in Lecture notes in computer science. , Cham: Springer Nature Switzerland, 2023, pp. 213–228.
[201] A. Kim, H. Kum, O. Roh, S. You, and S. Lee, "Robot gesture and user acceptance of information in human-robot interaction," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, in HRI '12. New York, NY, USA: Association for Computing Machinery, 2012, pp. 279–280. doi: 10.1145/2157689.2157793.
[202] D. Huang, D. G. Markovitch, and R. A. Stough, "Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust," *Journal of Retailing and Consumer Services*, vol. 76, p. 103600, 2024, doi: https://doi.org/10.1016/j.jretconser.2023.103600.
[203] C. Ferraro, V. Demsar, S. Sands, M. Restrepo, and C. Campbell, "The paradoxes of generative AI-enabled customer service: A guide for managers," *Bus Horiz*, vol. 67, no. 5, pp. 549–559, 2024, doi: https://doi.org/10.1016/j.bushor.2024.04.013.
[204] A. Aggarwal, C. C. Tam, D. Wu, X. Li, and S. Qiao, "Artificial intelligence-based chatbots for promoting health behavioral changes: Systematic review," *J. Med. Internet Res.*, vol. 25, p. e40789, Feb. 2023.
[205] Y. and L. D. and S. T. and E. K. and L. B. Laymouna Moustafa and Ma, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *J Med Internet Res*, vol. 26, p. e56930, Jul. 2024, doi: 10.2196/56930.
[206] G. Park, J. Chung, and S. Lee, "Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model," *Curr. Psychol.*, pp. 1–11, Nov. 2022.
[207] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1–2, pp. 119–155, Jul. 2003, doi: 10.1016/S1071-5819(03)00018-1.
[208] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, in HRI '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 407–414. doi: 10.1145/2559636.2559663.
[209] L. Seitz, "Artificial empathy in healthcare chatbots: Does it feel authentic?," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100067, 2024, doi: https://doi.org/10.1016/j.chbah.2024.100067.
[210] G. Hofree, B. A. Urgen, P. Winkielman, and A. P. Saygin, "Observation and imitation of actions performed by humans, androids, and robots: an EMG study," *Front. Hum. Neurosci.*, vol. 9, p. 364, Jun. 2015.

[211] run-llama, “GitHub LlamaIndex.” Accessed: Aug. 18, 2025. [Online]. Available: https://github.com/run-llama/llama_index
[212] Snowflake Inc., “Streamlit.” Accessed: Aug. 18, 2025. [Online]. Available: https://streamlit.io/
[213] SoftBank Robotics Europe, “NAOqi.” Accessed: Aug. 18, 2025. [Online]. Available: http://doc.aldebaran.com/2-5/index_dev_guide.html
[214] The PostgreSQL Global Development Group, “Postgres.” Accessed: Sep. 07, 2025. [Online]. Available: https://www.postgresql.org/
[215] RASA Technology Inc., “RASA.” Accessed: Aug. 18, 2025. [Online]. Available: https://rasa.com/
[216] Docker Inc., “Docker.” Accessed: Aug. 18, 2025. [Online]. Available: https://www.docker.com/
[217] nlmatics, “NLM Ingestor.” Accessed: Aug. 18, 2025. [Online]. Available: https://github.com/nlmatics/nlm-ingestor
[218] tiangolo, “FastAPI.” Accessed: Aug. 18, 2025. [Online]. Available: https://fastapi.tiangolo.com/
[219] M. Heerink, B. Krose, V. Evers, and B. Wielinga, “Influence of Social Presence on Acceptance of an Assistive Social Robot and Screen Agent by Elderly Users,” *Advanced Robotics*, vol. 23, pp. 1909–1923, Aug. 2009, doi: 10.1163/016918609X12518783330289.
[220] D. K. Lee, J. In, and S. Lee, “Standard deviation and standard error of the mean,” *Korean J. Anesthesiol.*, vol. 68, no. 3, pp. 220–223, Jun. 2015.
[221] J. M. Bland and D. G. Altman, “Statistics notes: Cronbach’s alpha,” *BMJ*, vol. 314, no. 7080, p. 572, Feb. 1997.
[222] K. Pearson, “X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material,” *Philos. Trans. R. Soc. Lond. A*, vol. 186, no. 0, pp. 343–414, Dec. 1895.
[223] K. Pearson, “``DAS FEHLERGESETZ UND SEINE VERALLGEMEINER-UNGEN DURCH FECHNER UND PEARSON.’’ A REJOINDER,” *Biometrika*, vol. 4, no. 1–2, pp. 169–212, Jun. 1905.
[224] F. Suhail, M. Adel, M. Al-Emran, and A. A. AlQudah, “Are students ready for robots in higher education? Examining the adoption of robots by integrating UTAUT2 and TTF using a hybrid SEM-ANN approach,” *Technol Soc*, vol. 77, p. 102524, 2024, doi: https://doi.org/10.1016/j.techsoc.2024.102524.
[225] H. Yildiz Durak and A. Onan, “Predicting the use of chatbot systems in education: a comparative approach using PLS-SEM and machine learning algorithms,” *Curr. Psychol.*, vol. 43, no. 28, pp. 23656–23674, Jul. 2024.
[226] N. Fragakis, G. Trichopoulos, and G. Caridakis, “Empowering education with intelligent systems: Exploring large language models and the NAO robot for information retrieval,” *Electronics (Basel)*, vol. 14, no. 6, p. 1210, Mar. 2025.
[227] C. Y. Kim, C. P. Lee, and B. Mutlu, “Understanding large-language model (LLM)-powered human-robot interaction,” 2024.

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Formelverzeichnis

11 Abkürzungsverzeichnis

Adaptive Learning-Technologie ..... *ALT*
Allgemeine Fragen ..... *AF*
Application-Programming -Interface ..... *API*
Approximate Nearest Neighbor ..... *ANN*
arithmetisches Mittel ..... $\bar{x}$
Artificial Intelligence in Education ..... *AIEd*
Aufwandserwartung ..... *AE*
Autismus-Spektrum-Störung ..... *ASD*
Autismus-Spektrum-Störungen ..... *ASS*
chain-of-thought ..... *COT*
Conversational AI ..... *CAI*
Cronbach's Alpha ..... α
Cross-Encoder-Reranker ..... *CER*
Datenschutzgrundverordnung ..... *DSGVO*
Deep Learning ..... *DL*
Degrees of Freedom ..... *DOF*
Durchschnittswert ..... $\bar{x}$
Embedding-Modelle ..... *EM*
Erleichternde Bedingungen ..... *EB*
Generative KI-Systeme ..... *GKI*
Gewohnheit ..... *GH*
Hedonistische Motivation ..... *HM*
Historically Black Colleges and Universities ..... HBCU
Human-Centred Learning Analytics ..... *HCLA*
Human-Computer Interaction ..... *HCI*
Human-in-the-Loop ..... HITL
Innovation Diffusion Theory ..... *IDT*
intelligente tutorielle Systeme ..... *ITS*
Internet of Things ..... *IoT*
Kindergarten bis zur 12. Klasse *K-12*
Knowledge Graph Embedding ..... *KGE*
Künstlicher Generativer Intelligenz ..... *KGI*
Künstlicher Intelligenz ..... KI
Large Language Modelle ..... *LLM*
Leistungserwartung ..... *LE*
Machine Learning ..... *ML*
Mathematik, Informatik, Naturwissenschaft und Technik ..... *MINT*
Maximalwert ..... *max*
Mensch-Maschine-Kommunikation ..... MMK
Mensch-Roboter-Interaktion ..... MRI
Minimalwert ..... *min*
Mittelwert ..... $\bar{x}$
Mix-of-Granularity ..... *MoG*
Natural Language Model ..... *NLM*
Natural Language Processing ..... *NLP*
Natural Language Understanding ..... *NLU*
objektrelationales Datenbankmanagementsystem ..... ORDBMS
okal datenschutzfreundlichen RAG-Ansatz ..... *LPRAG*
Open-Source-Framework ..... *OSF*
Optical Character Recognition ..... *OCR*
Pair Programming ..... *PP*
Perceived Ease of Use ..... *PEOU*

Perceived Usefulness ..... *PU*
Physical Human-Robot Interaction ..... pHRI
Post-Test ..... *POT*
Preis Wert ..... *PW*
Pre-Test ..... *PRT*
Programmed Logic for Automatic Teaching Operations ..... *PLATO*
Retrieval Augmented Generation ..... RAG
selbstregulierten Lernens ..... SRL
Self-Attention-Mechanismen ..... *SAM*
Small Language Models ..... *SLM*
Software Development Kit ..... *SDK*
soziale Einflüsse ..... *SE*
Sozialer Einfluss ..... *SE*
Standardabweichung ..... σ
Sustainable Development Goals ..... *SDG*
System Usability Scale ..... *SUS*
Technology Acceptance Model ..... TAM
Technology–Organization–Environment ..... *TOE*
Theory of Mind ..... *ToM*
Theory of Planned Behavior ..... *TPB*
Theory of Reasoned Action ..... *TRA*
Unified Theory of Acceptance and Use of Technology ..... *UTAUT*
Verhaltensabsichten ..... *VA*

12 Anhang

12.1 UTAUT – deskriptive Statistik nach Fragen

Frage	min	x̄	max	σ
Leistungserwartung				
LE1	2,00	3,80	5,00	1,03
LE2	1,00	3,70	5,00	1,25
LE3	1,00	3,40	5,00	1,17
LE4	1,00	2,70	4,00	0,95
Aufwandserwartung				
AE1	2,00	4,00	5,00	0,94
AE2	3,00	4,10	5,00	0,57
AE3	3,00	3,90	5,00	0,88
AE4	3,00	3,80	5,00	0,79
Sozialer Einfluss				
SE1	1,00	2,80	4,00	1,14
SE2	2,00	2,70	3,00	0,48
SE3	3,00	3,50	4,00	0,53
SE4	3,00	4,30	5,00	0,67
erleichternde Bedingungen				
EB1	1,00	3,90	5,00	1,45
EB2	2,00	3,90	5,00	0,99
EB3	1,00	4,10	5,00	1,20
Verhaltensabsicht				
VA1	1,00	3,20	4,00	1,03

VA2	1,00	3,50	5,00	1,18
VA3	1,00	3,00	5,00	1,25

Tabelle 43 - Deskripitive Statistik - NAO (UTAUT)

Frage	**min**	**x̄**	**max**	**σ**
Leistungserwartung				
LE1	2,00	3,80	5,00	0,79
LE2	2,00	3,60	5,00	0,97
LE3	1,00	3,40	5,00	1,07
LE4	2,00	3,40	5,00	0,97
Aufwandserwartung				
AE1	2,00	3,70	5,00	1,25
AE2	2,00	4,20	5,00	1,23
AE3	2,00	4,30	5,00	1,06
AE4	2,00	3,80	5,00	1,23
Sozialer Einfluss				
SE1	2,00	2,80	4,00	0,79
SE2	2,00	3,10	5,00	0,99
SE3	2,00	3,40	5,00	0.97
SE4	2,00	4,10	5,00	0,99
erleichternde Bedingungen				
EB1	3,00	4,30	5,00	0,67
EB2	3,00	4,30	5,00	0,82
EB3	2,00	3,80	5,00	0,92
Verhaltensabsicht				
VA1	1,00	3,10	4,00	1,29
VA2	2,00	3,80	5,00	0,92

VA3	1,00	3,40	5,00	1,26

Tabelle 44 - Deskripitive Statistik - Webchatbot (UTAUT)

12.2 System Prompts für die RAG Pipeline

Eine RAG-Pipeline benötigt immer ein Prompt-Template, um die abgerufenen Informationen korrekt in das Sprachmodell einzubinden.
Das Template definiert die Struktur, in der Kontextdaten, Nutzerfrage und mögliche Zusatzhinweise zusammengeführt werden. Ohne ein solches Template könnte das Modell die abgerufenen Dokumente nicht konsistent in die Antwortgenerierung integrieren. Darüber hinaus sorgt das Template für Wiederholbarkeit und Vergleichbarkeit der Ergebnisse. Es bildet somit die zentrale Schnittstelle zwischen der Retrieval-Komponente und der Generierungsphase der RAG-Pipeline.

12.2.1 Answer Mode

Du bist **GenAI-Mentor**, ein freundlicher, motivierender Lernassistent für Studierende im Fach *Generative Artificial Intelligence* (Sprach- & Bildmodelle, Diffusion, RLHF, Prompt-Engineering …).
- Aufgaben:
1. Erkläre Fachbegriffe und Konzepte in klarer, präziser Sprache.
2. Gib praxisnahe Beispiele, Code-Snippets oder Analogien, um das Verständnis zu vertiefen.
3. Füge wenn sinnvoll kurze Diagrammbeschreibungen oder Schritt-für-Schritt-Abläufe hinzu (z. B. „So funktioniert ein Diffusion-Sampler“).
- Regeln:
- Antworte **ausschließlich** mit Informationen aus den bereitgestellten Kontextpassagen.
- Sprich die Studierenden direkt mit „du“ an.
- Verwende keine Halluzinationen und keine externen Links, die nicht im Kontext stehen.

{context_str}
\n---\n
Question:\n
{query_str}\n\n

12.2.2 Question Mode

Du bist **Quiz-Coach**, ein interaktiver Lernassistent für das Fach *Generative Artificial Intelligence*.
- Deine Hauptaufgabe: Stelle dem Studierenden sinnvolle Fragen, um sein Verständnis schrittweise zu prüfen und zu vertiefen.
- Vorgehensweise:
1. Formuliere **eine klar abgegrenzte Frage** zu den Konzepten aus dem Kontext.
- Beginne mit leichten Verständnisfragen und erhöhe bei korrekten Antworten langsam den

Schwierigkeitsgrad (Bloom-Taxonomie: Remember → Understand → Apply → Analyze → Evaluate → Create).
2. Fordere den Studierenden explizit zur Antwort auf.

Beispiel-Einleitung: *„Deine nächste Aufgabe:“* / *„Was meinst du?“*

3. Warte anschließend auf die Antwort, gib **keine** vollständige Lösung preis.
- Antworte nur mit kurzen Bestärkern wie „Okay, ich warte auf deine Antwort …“.

---------\n

{context_str}
\n---------\n
Question:\n
{query_str}\n\n