
Der Dialog-Trainer

Erprobung eines Prompts für Selbstlernphasen im Sprachunterricht
in zwei generativen Chatbots

Autoren

Thomas Schatz

Alisa Poleshchuk

Sabrina Becker

Hinweis zum Datenschutz

Mitarbeiter und Mitarbeiterinnen der Hochschule Harz sind dringend angehalten, den Einsatz des jeweiligen KI-Programms eigenständig auf datenschutzrechtliche Zulässigkeit hin zu überprüfen. Sofern möglich, sollten Verarbeitungen lokal, d.h. ohne Übertragung von (personenbezogenen) Daten an externe Dienstleister. In der Regel wird die Übertragung von personenbezogenen Daten (bspw. Namen, Matrikelnummern, Notenlisten, Prüfungsarbeiten) an externe Dienstleister und Unternehmen unzulässig sein, von einer Eingabe personenbezogener Daten ist deswegen grundsätzlich abzugehen. Dies gilt auch und insbesondere für interne Dokumente (bspw. Briefe, Bescheide etc.), deren Nutzung nur hochschulintern vorgesehen ist. Vorsicht ist bei vorinstallierten Programmen oder neu hinzugefügten Funktionen geboten, weil sich auch hinter diesen Assistenzsystemen (externe) KI-Programme verbergen können.

Wenn Sie unsicher sind, ob das von Ihnen benutzte Programm oder eine Browsererweiterung unzulässig Daten überträgt, sollten Sie vom ungeprüften Einsatz der Software Abstand nehmen oder Rücksprache mit dem Datenschutzbeauftragten halten. Die Nutzung von KI-Programmen wird regelmäßig zulässig sein, sofern keine personenbezogenen oder sonstige schützenswerte Daten verwendet werden.

Inhaltsverzeichnis

1	Beschreibung des Vorhabens	4
2	Entwicklung und Erprobung des Prompts	5
3	Der Prompt	6
4	Chatverlauf	8
5	Diskussion und Fazit	16

Tabellenverzeichnis

Tab. 1:	Bedeutung und Anwendung der Markdown-Symbole	6
Tab. 2:	Auftaktfrage und -antworten der Chats	9
Tab. 3:	Zweite Fragen und Antworten.....	10
Tab. 4:	Dritte Fragen und Antworten.....	11
Tab. 5:	Vierte Fragen und Antworten	12
Tab. 6:	Fünfte Fragen und Antworten	13
Tab. 7:	Sechste Frage und Antwort	14
Tab. 8:	Zusammenfassendes Resümee der Chatbots.....	15

1 Beschreibung des Vorhabens

Das TeachingLab der Hochschule Harz ist eine dem Rektorat zugeordnete Organisationseinheit, deren Aufgabe seit dem Jahr 2019 darin besteht, den Lehrkörper beim Einsatz digitaler Elemente in Lehre und Prüfungen zu unterstützen. Die Einrichtung des TeachingLabs ist eine Umsetzungsmaßnahme der ebenfalls 2019 beschlossenen Digitalisierungsstrategie der Hochschule.

Seit Beginn des Jahres 2023 ist auch an unserer Hochschule die Frage nach den Auswirkungen generativer Chatbots auf Lehre und Prüfungsmodalitäten relevant. Während ein Teil des Kollegiums nach wie vor zurückhaltend auf KI reagiert, erkennt eine wachsende Anzahl der Lehrenden das Potenzial zur Bereicherung der synchronen und asynchronen Lehre, zur Optimierung von Zeit und Flexibilität durch Delegation von Routineaufgaben an maschinelle Systeme sowie zur Vermittlung von KI-Kompetenzen an Studierende. Das TeachingLab unterstützt diese Entwicklung durch ein breites Spektrum an Informations- und Weiterbildungsangeboten sowie durch individuelle Beratung bei spezifischen Projekten.

So entstand auch der vorliegende Text durch einen Impuls der Leiterinnen des Sprachzentrums unserer Hochschule, die das TeachingLab baten, im Rahmen einer Dienstberatung Anwendungsszenarien für generative Chatbots in der Vermittlung von Sprachkompetenzen vorzustellen. In Kooperation mit Kollegen des Verbundprojektes eSALSA wurden konkrete Einsatzmöglichkeiten konzipiert, evaluiert und dem Kollegium des Sprachzentrums präsentiert.

Eines der in diesem Kontext entwickelten didaktischen Szenarien setzt auf dem sogenannten "Dialog-Trainer" auf. Dieser im TeachingLab entwickelten Prompt soll in Selbstlernphasen des Sprachunterrichts eine Möglichkeit schaffen, die schriftlichen Sprachkompetenzen der Lernenden in praxisnahen Dialogen zu stärken, wobei die KI-Technologie die Rollen eines Dialogpartners und eines Lehrenden übernimmt. Lehrkräfte können durch Variation des Prompts den Lernprozess gezielt steuern und gleichzeitig abwechslungsreich und motivierend gestalten. Wert wurde daraufgelegt, durch unmittelbares Feedback des KI-Systems auf die Eingaben der Studierenden eine unmittelbare Fehlererkennung und -korrektur zu ermöglichen und so die Effizienz des Selbstlernprozesses zu steigern.

Die praktische Anwendung in Lehrveranstaltungen könnte sich folgendermaßen gestalten: Die Lehrperson adaptiert die Parameter des Prompt-Beispiels, wie etwa Sprache, Kompetenzniveau, Dialogthema und Rollenverteilung, an die spezifische Lehr- und Lernsituation. Anschließend erhalten die Lernenden den modifizierten Prompt als Kopiervorlage. Sie fügen den Prompt in die Befehlszeile des verwendeten KI-Tools ein und führen die Dialogübung durch. Zur Qualitätssicherung besteht die Option, den generierten Dialog in einem Textverarbeitungsprogramm zu speichern und der Lehrkraft zur Verfügung zu stellen.

Der vorliegende Text dokumentiert den Prozess der Überprüfung zweier Hypothesen, die während der Erprobungsphase des Prompts im Team des TeachingLabs und der Mitarbeiter des Projektes eSALSA diskutiert wurden. Die erste Hypothese behauptet, dass eine Evaluation des Prompts in verschiedenen KI-Tools lohnenswert ist, da die Qualitätsunterschiede in der Aufgabenerfüllung erheblich sein könnten. Setzt man dies voraus, ist eine zweite Hypothese die konsequente Fortsetzung: In der Vermittlung von Sprachkompetenzen können generative Chatbots angewendet werden, die im Vergleich mit dem etablierten ChatGPT 4o zumindest konkurrenzfähig sind und in den didaktischen Konzepten von Lehrenden deshalb eine Rolle spielen sollten.

In folgenden Abschnitten präsentieren wir in komprimierter Form das Design zur Erprobung des Prompts und der Überprüfung der genannten Hypothesen. Zunächst stellen wir überblicksartig den Prozess des Prompt-Engineerings dar und erläutern das Untersuchungsdesign. Der dritte Teil des Papers ist der Präsentation des Prompts und Hinweisen auf einzelne Details seiner Gestaltung vorbehalten. Daran anschließend werden die Chatverläufe dokumentiert, die sich durch die Anwendung der Prompts im Dialog mit zwei KI-Tools entwickelten. Der Text schließt mit einer analytischen Interpretation der Ergebnisse.

2 Entwicklung und Erprobung des Prompts

Der "Dialog-Trainer"-Prompt wurde im Juli 2024 im TeachingLab der Hochschule Harz entwickelt und evaluiert. Initial als Customer-GPT konzipiert, ergab die Analyse, dass die praktischen Limitationen die Vorteile der Konstruktion überwogen. Zwei Hauptgründe waren ausschlaggebend: Erstens die Notwendigkeit eines kostenpflichtigen ChatGPT-Abonnements für die Lernenden, die den „Dialog-Trainer“ würden nutzen wollen, was nicht als gegeben vorausgesetzt werden konnte. Zweitens die Restriktion auf ChatGPT, wodurch potenzielle Qualitätsvorteile alternativer Chatbots nicht genutzt werden konnten.

Infolgedessen entschied sich das Forschungsteam für eine Copy-Paste-Variante. Diese Methode ermöglicht es Lehrenden, den Prompt als flexibles und leicht adaptierbares Template zur Verfügung zu stellen, welches sie dann, verbunden mit spezifischen Aufgabenstellungen, an die Lernenden weitergeben können.

Nach der Entwicklung des Prompts erfolgte eine Erprobung in zwei Chatbots: ChatGPT 4.0, zugänglich für alle Angehörigen der Hochschule Harz über die HAWKi-API, und Claude.ai, welcher eine persönliche Anmeldung im browserbasierten Tool erforderte. Diese Testphase diente zugleich der Validierung der Hypothesen. Der Prompt wurde in der im folgenden Abschnitt dokumentierten Textversion verwendet. Als Testperson fungierte ein Mitarbeiter des TeachingLabs mit basalen Englischkenntnissen. Seine Aufgabe bestand darin, in der Rolle des Kunden einen Dialog in englischer Sprache auf Niveau A2 des Gemeinsamen Europäischen Referenzrahmens mit einer Verkäuferin in einem Brautmodengeschäft zu führen.

Der Dialog war auf fünf Frage-Antwort-Sequenzen konzipiert, wobei die KI die Rolle der Verkäuferin übernahm. Zusätzlich hatte die KI die Aufgabe, analog zu einem Englischlehrer, die Antworten der Testperson anhand von drei Leitfragen zu evaluieren, auf Fehler hinzuweisen und Korrekturvorschläge zu unterbreiten sowie Ideen zur Verbesserung der Ausdrucksweise vorzuschlagen. Abschließend sollte die KI ein zusammenfassendes Resümee ziehen.

Das Monitoring wurde von einer Co-Autorin durchgeführt, die als wissenschaftlichen Mitarbeiterin im Verbundprojekt eSALSA tätig ist und über einen Masterabschluss als Lehrerin für das Fach Deutsch als Fremdsprache verfügt.

Gemäß der Prompt-Programmierung begannen die Dialoge in beiden Chatbots mit derselben Frage. Eine vollständige Synchronität in den folgenden Sequenzen konnte nicht erreicht werden. Der Fokus der Untersuchung lag jedoch primär auf der Qualität der KI-Rückmeldungen zu den Antwort-Eingaben der Testperson, die als Äquivalent zum Feedback eines Englischlehrers konzipiert waren. Die von den KI-Tools generierten Ergebnisse wurden anschließend im Forschungsteam vor dem Hintergrund der formulierten Hypothesen analysiert und diskutiert.

3 Der Prompt

Für den „Dialog-Trainer“ wurde in mehreren Prozessschritten ein relativ umfassender Prompt geschrieben. In der ersten Phase der Annäherung an ein funktionierendes Ergebnis diskutierte das Team das Einsatzfeld und die notwendigen inhaltlichen Elemente eines „Dialog-Trainers“. Es wurde entschieden, dass der Prompt in Selbstlernphasen zum Einsatz kommen solle, den Nutzerinnen und Nutzern aber trotz der Abwesenheit einer Lehrperson Korrektur- und Feedbackfunktionen bieten solle.

Davon ausgehend wurden im zweiten Schritt die Kontextinformationen, Rollenzuweisungen, die Aufgabe und sowie die Anweisungen für die Teilschritte in den Dialog- und Feedbackschleifen formuliert. Dazu nutzten die Teammitglieder eine Textverarbeitungssoftware, um kollaborativ an der Präzisierung der Formulierungen arbeiten zu können. Parallel dazu wurde der Prompt in ChatGPT 4o fortlaufend getestet, um Hinweise auf Unzulänglichkeiten und Verbesserungspotenziale zu generieren.

In Übereinstimmung mit Empfehlungen zum Prompt-Engineering wurden die Markdown-Symbole "#", "##" und "*" verwendet, um die Verständlichkeit des Prompt-Textes für KI-Tools zu maximieren. Solche Sonderzeichen spielen eine zentrale Rolle bei der Strukturierung und Betonung von Inhalten, was sowohl für das menschliche Verstehen als auch die maschinelle Verarbeitung von großer Bedeutung ist. Markdown-Symbole helfen, den Text systematisch zu gliedern und erleichtern es KI-Anwendungen, den logischen Zusammenhang zwischen einzelnen Textpassagen des Prompts korrekt zu interpretieren. In der folgenden Tabelle wird die Bedeutung der verwendeten Symbole kurz erklärt.

Tab. 1: Bedeutung und Anwendung der Markdown-Symbole

Symbol	Bedeutung und Anwendung
#	<ul style="list-style-type: none"> • Das einfache Raute-Symbol wird verwendet, um eine Überschrift der ersten Ebene (Hauptüberschrift) zu erstellen. • Im Prompt wurde das Raute-Symbol genutzt, um die Platzierung der wesentlichsten Informationen „Kontext“, „Rollenverteilung“, „Sprachniveau“ und „Deine Aufgabe“ herauszuheben und gegeneinander abzugrenzen • Zwischen Raute-Symbol und folgendem Text ist ein Leerzeichen zu setzen. • Zusätzlich wurden im Prompt vor und hinter Textzeilen mit dem Symbol Leerzeilen eingefügt.
##	<ul style="list-style-type: none"> • Das doppelte Raute-Symbol dient zur Erstellung einer Überschrift der zweiten Ebene (Teilüberschrift). • Im Prompt wurde das Raute-Symbol genutzt, um die einzelnen Prozessschritte innerhalb des gewünschten Ablaufs von Dialog- und Feedbackschleifen kenntlich zu machen und abzugrenzen. • Zwischen Raute-Symbol und folgendem Text ist ein Leerzeichen zu setzen. • Zusätzlich wurden im Prompt vor und hinter Textzeilen mit dem Symbol Leerzeilen eingefügt.
*	<ul style="list-style-type: none"> • Das Stern-Symbol wird verwendet, um eine Aufzählungsliste zu erstellen. • Ein Stern am Anfang einer Zeile markiert einen Listenpunkt, der eine Aufzählung oder eine Liste einleitet. • Dadurch kann die KI Informationen als separate, aber inhaltlich zusammenhängende Einheiten erkennen und verarbeiten.

Zu dem Zeitpunkt als „Dialog-Trainer“ zuverlässig zur Zufriedenheit der Teammitglieder funktionierte, lag er in folgender Form und Formulierung vor:

Kontext

Du bist Englisch-Lehrer an einer deutschen Hochschule. Dein Ziel ist es, deutschen Studenten die Fremdsprache Englisch durch Dialoge zu vermitteln. Der Dialog ist ein Verkaufsgespräch in einem Fachgeschäft für Brautmoden.

Rollenverteilung

- Du bist der Verkäufer.*
- Der Student ist der Kunde.*

Sprachniveau

A2

Deine Aufgabe

- * Beginne mit Schritt 1.*
- * Führe Schritt 3 nur aus, wenn der Kunde geantwortet hat.*
- * Wenn Du Schritt 3 ausgeführt hast, führst Du die weiteren Schritte nacheinander aus.*

Schritt 1

Schreibe: „Hello, welcome in our shop. What can I do for you?“

Schritt 2

- * Warte die Antwort des Studenten ab.*
- * Vermeide es, die folgenden Schritte in deiner Antwort aufzuzählen, führe sie einfach nur aus.*

Schritt 3

Analysiere die Antwort und achte auf folgende Leitfragen der Text-Analyse:

- * „Wurden die passenden Vokabeln genutzt?“*
- * „Sind Grammatik und Rechtschreibung korrekt?“*
- * „Ist der Ausdruck gut und verständlich?“*

Schritt 4

- * Gib nach der Analyse das kurze Feedback eines Englisch-Lehrers zu der Antwort.*
- * Gehe in dem Feedback einzeln auf die Leitfragen der Text-Analyse ein.*
- * Gib das Feedback immer auf Deutsch.*

Schritt 5

Wenn die Antwort Fehler enthält, weißt Du nacheinander auf die Fehler hin.

Schritt 6

Zeige, wie es in der englischen Sprache richtig wäre.

Schritt 7

Zeige eine Antwort, die aus deiner Sicht auf dem Sprachniveau A2 perfekt wäre.

Schritt 8

Stelle die nächste Frage und gehe den Ablauf erneut durch.

Schritt 9

Stelle insgesamt fünf Fragen, bei denen du den Ablauf durchgehst.

Schritt 10

Wenn die fünfte Frage abgearbeitet ist, ziehe ein zusammenfassendes Resümee. Dieses Resümee sollte der abschließenden Zusammenfassung eines Englisch-Lehrers entsprechen. Dieses Resümee sollte eine Beurteilung der Leistung und Hinweise darauf enthalten, was der Student gezielt üben sollte, um sich zu verbessern.

Schritt 11

Beende den Dialog abschließend mit dem Satz: "Ich hoffe, es hat Dir Spaß gemacht und Du konntest etwas lernen."

Der Prompt und der Dialog im Brautmodengeschäft dienen als Modell, um die Funktionalität des "Dialog-Trainers" umfassend zu testen. In der späteren Verwendung können Lehrende den Prompt durch Variation von Parametern wie der Sprache, dem Sprachniveau, der Dialogsituation oder der Anzahl der Frage-Antwort-Schleifen beliebig verändern. Somit wäre möglich, den Prompt auf die Erfordernisse der jeweiligen Lehr- und Lernsituationen anzupassen.

Darüber hinaus könnte der "Dialog-Trainer" auch sehr gut genutzt werden, um sich individuell auf spezifische Situationen vorzubereiten und kontextualisierende Vokabeln und Redewendungen gezielt zu trainieren. Dies könnte besonders hilfreich sein, um in realen Gesprächen sicherer und flüssiger zu kommunizieren.

4 Chatverlauf

Auf den folgenden Seiten werden die Protokolle der Chats präsentiert, die nach der Entwicklung und Implementierung des „Dialog-Trainer“-Promptes in den generativen Chatbots ChatGPT 4o (Zugang via HAWKi-API) und Claude.ai (hinfort: Claude) entstanden sind. Diese Dialoge hatten nicht mehr das Ziel, den „Dialog-Trainer“-Prompt weiter zu optimieren. Vielmehr zielte dieser Durchlauf darauf ab, einen Beitrag zur Beantwortung der Hypothesen zu leisten und zu zeigen, welche Ergebnisse mit dem besagten Prompt in unterschiedlichen Chatbots erzielt werden können.

In den nachfolgenden Tabellen sind die beiden Chats in strukturierter Form wiedergegeben. Diese tabellarische Darstellung ermöglicht es, die in der Reihenfolge der Fragen aufgeführten Dialogsequenzen zu isolieren und direkt gegenüberzustellen. Aus unserer Sicht erleichtert dieses methodische Vorgehen die Vergleichbarkeit der erzielten Resultate erheblich.

Im Anschluss an die Tabellen folgt eine beschreibende Auflistung der Beobachtungen und Feststellungen, die von den Teammitgliedern in Bezug auf die einzelnen Dialogsequenzen gemacht wurden. Diese Analyse bereitet die, dem abschließenden Kapitel vorbehaltene Bewertung der Ergebnisse vor und unterstützt die Ableitung fundierter Schlussfolgerungen zum Einsatz des „Dialog-Trainers“.

Tab. 2: Auftaktfrage und -antworten der Chats

Initiale Frage der KI: Hello, welcome in our shop. What can I do for you?	
Antwort der Testperson Hello, I'm looking for a simple wedding dress.	
ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Ja, Grammatik und Rechtschreibung sind korrekt.</p> <p>Ist der Ausdruck gut und verständlich? Ja, der Ausdruck ist verständlich und klar.</p> <p>Feedback Sehr gut gemacht! Deine Antwort verwendet treffende Vokabeln und ist grammatikalisch korrekt. Der Ausdruck ist klar und verständlich.</p>	<p>Analyse:</p> <ul style="list-style-type: none"> • Passende Vokabeln wurden genutzt: "wedding dress" ist korrekt. • Grammatik und Rechtschreibung sind korrekt. • Der Ausdruck ist gut und verständlich. <p>Feedback: Gut gemacht! Du hast einen klaren und korrekten Satz formuliert. Die Verwendung von "simple" zeigt, dass du spezifisch ausdrücken kannst, was du suchst.</p> <p>Eine perfekte Antwort auf A2-Niveau könnte so aussehen: "Hello, I'm looking for a simple wedding dress. Can you help me?"</p>

Beobachtungen und Feststellungen

Die beiden Chatbots stimmen in ihrem Urteil überein und bewerten die eingegebene, sehr einfache Antwort der Testperson als in Ordnung. Dabei zeigen sich jedoch Unterschiede in der methodischen Vorgehensweise der beiden Systeme. ChatGPT geht formaler vor, indem jeweils die im Prompt verankerten Leitfragen genannt werden, bevor die zugehörigen Antworten folgen. Claude hingegen verzichtet auf die Nennung der Leitfragen und fasst die Analyse der Antwort in Stichpunkten zusammen, wobei jeweils ein Stichpunkt einer Leitfrage zugeordnet ist.

Das generierte Feedback beider Chatbots wirkt aussagekräftig und zufriedenstellend. Es fällt jedoch auf, dass ChatGPT die im Prompt ausdrücklich formulierte Aufforderung, „Zeige eine Antwort, die aus deiner Sicht auf dem Sprachniveau A2 perfekt wäre“, nicht berücksichtigt. Eine plausible Erklärung für dieses Versäumnis könnte darin liegen, dass ChatGPT aufgrund der fehlenden Diagnose eines Fehlers in der Antwort der Testperson keine Veranlassung sah, ein perfektes A2-Niveau-Beispiel zu präsentieren.

Claude hingegen erfüllt die Vorgabe aus dem Prompt vollständig. Das bedeutet, dass Claude auch dann, wenn keine Fehler in der Antwort der Testperson diagnostiziert wurden, die geforderte beispielhafte Antwort auf dem Sprachniveau A2 präsentiert. Diese vollständige Abarbeitung des Prompts durch Claude zeigt eine konsequentere Umsetzung der Anforderungen.

Tab. 3: Zweite Fragen und Antworten

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Frage 2 Do you have a specific style in mind?</p>	<p>Frage 2: Of course! What size are you looking for?</p>
<p>Antwort der Testperson Oh, I had a dream. I would dress me in a suit made of gold.</p>	<p>Antwort der Testperson Sorry, i don't know my size. Look at me, my legs are short and my body is fully. Which size would you prefer?</p>
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind größtenteils passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Nicht ganz; es gibt einen Grammatikfehler.</p> <p>Ist der Ausdruck gut und verständlich? Der Ausdruck ist größtenteils verständlich, aber der Grammatikfehler könnte verwirrend sein.</p> <p>Feedback Deine Antwort ist größtenteils verständlich und die Vokabeln sind passend. Es gibt jedoch einen kleinen Grammatikfehler.</p> <p>Fehlerhinweis Der Satz "I would dress me in a suit made of gold" ist nicht ganz korrekt.</p> <p>Korrektur Es sollte heißen: "I would dress in a suit made of gold."</p> <p>Perfekte Antwort auf A2-Niveau "Oh, I had a dream. I would dress in a suit made of gold."</p>	<p>Analyse:</p> <ul style="list-style-type: none"> • Vokabeln: Größtenteils passend, aber einige Fehler ("fully" statt "full"). • Grammatik: Kleinere Fehler, z.B. "i" statt "I". • Ausdruck: Verständlich, aber mit Verbesserungspotenzial. <p>Feedback: Deine Antwort ist grundsätzlich verständlich, aber es gibt einige Punkte, die verbessert werden können:</p> <ol style="list-style-type: none"> 1. "i" sollte großgeschrieben werden: "I". 2. "fully" ist nicht das richtige Wort für einen kräftigen Körperbau. "full" wäre hier passender. 3. Der Satz "Look at me" ist etwas abrupt. Eine höflichere Formulierung wäre besser. 4. "prefer" passt nicht ganz in diesem Kontext. "recommend" wäre besser. <p>Korrekte Version: "Sorry, I don't know my size. As you can see, my legs are short and my body is full. Which size would you recommend?"</p> <p>Eine perfekte Antwort auf A2-Niveau könnte so aussehen: "I'm sorry, I don't know my size. I have short legs and a full body. Can you help me find the right size?"</p>

Beobachtungen und Feststellungen

Mit den zweiten Fragen trennen sich die Dialoge: Die Anschlussfragen der Chatbots auf die erste Antwort der Testperson unterscheiden sich, ebenso wie die darauffolgenden Antworten. Dabei zeigen sich signifikante Unterschiede im Umfang der Eingaben: Während die Testperson dem Chatbot Claude eine Antwort mit 87 Zeichen zur Verfügung stellt, umfasst die Antwort an ChatGPT lediglich 45 Zeichen. Diese Diskrepanz spiegelt sich dann auch in den Längen des Feedbacks wider, wobei die Rückmeldung des Chatbots Claude aufgrund der höheren Anzahl an Fehlern entsprechend ausführlicher ausfällt.

Beide Chatbots sind in der Lage, Fehler in den Antworten der Testperson zu erkennen und darauf abgestimmte Korrekturvorschläge zu unterbreiten. Die Art und Weise, wie diese Fehler und Korrekturvorschläge präsentiert werden, variiert jedoch. Beide Varianten werden aber als gut nachvollziehbar empfunden.

Es wird weiterhin festgestellt, dass Claude in dieser Sequenz nicht vollständig korrekt korrigiert. In einem spezifischen Fall, in dem das Wort „fully“ korrigiert werden sollte, wäre wohl „curvy“ dem Verbesserungsvorschlag „full“ vorzuziehen gewesen.

Es fällt auf, dass sich ChatGPT mit seinem Feedback sehr eng an die ursprünglichen Antwort der Testperson anlehnt. Es liefert präzise Rückmeldungen, die sich direkt auf die eingegebene Antwort beziehen. Darüber hinaus gehende Hinweise unterbleiben.

Claude hingegen nutzt die Prompt-Anweisung „Eine perfekte Antwort auf A2-Niveau könnte so aussehen:“ um weitergehende Vorschläge anzubieten. Die Antwort der Testperson wird dabei von der KI variantenreich reformuliert und bietet eine breitere Perspektive auf mögliche Verbesserungen. Dieser Befund gilt auch für alle folgenden Dialogsequenzen.

Tab. 4: Dritte Fragen und Antworten

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Nächste Frage We don't have a suit made of gold, but we have many beautiful dresses. Do you like long dresses or short dresses?</p>	<p>Nächste Frage: No problem, I can help you find the right size. Let's try on a few dresses. Do you prefer a long or short dress?</p>
<p>Antwort der Testperson I like the color red. I would dress in red suit with short sleeves and long trousers.</p>	<p>Antwort der Testperson I like the color red. I would dress in red suit with short sleeves and long trousers.</p>
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind größtenteils passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Es gibt einige kleine Fehler.</p> <p>Ist der Ausdruck gut und verständlich? Der Ausdruck ist größtenteils verständlich, aber die kleinen Fehler könnten verwirrend sein.</p> <p>Feedback Deine Antwort ist größtenteils verständlich und die Vokabeln sind gut. Es gibt aber einige kleine Grammatikfehler.</p> <p>Fehlerhinweis Der Satz "I would dress in red suit with short sleeves and long trousers" ist nicht ganz korrekt. Es fehlt ein Artikel vor "red suit".</p> <p>Korrektur Es sollte heißen: "I would dress in a red suit with short sleeves and long trousers."</p> <p>Perfekte Antwort auf A2-Niveau "I like the color red. I would dress in a red suit with short sleeves and long trousers."</p>	<p>Analyse:</p> <ul style="list-style-type: none"> • Vokabeln: Größtenteils passend, aber einige Ungenauigkeiten. • Grammatik: Kleinere Fehler, z.B. fehlendes "a" vor "red suit". • Ausdruck: Verständlich, aber nicht ganz passend zur Frage. <p>Feedback: Deine Antwort zeigt, dass du deine Vorlieben ausdrücken kannst, aber es gibt einige Punkte zu verbessern:</p> <ol style="list-style-type: none"> 1. Du hast nicht direkt auf die Frage geantwortet, ob du ein langes oder kurzes Kleid bevorzugst. 2. Du sprichst von einem Anzug ("suit"), aber in einem Brautmodengeschäft geht es um Kleider. 3. Der Satz "I would dress in red suit" benötigt den unbestimmten Artikel "a". 4. Die Beschreibung der Kleidung passt nicht ganz zum Kontext eines Brautkleides. <p>Korrekte Version: "I like the color red. I would like a red dress with short sleeves and a long skirt."</p> <p>Eine perfekte Antwort auf A2-Niveau könnte so aussehen: "I prefer a long dress. Can I try on a red one? I like dresses with short sleeves."</p>

Beobachtungen und Feststellungen

Auf die Anschlussfragen der Chatbots antwortet die Testperson nun wieder mit identischen Sätzen. Beide Chatbots erkennen denselben Fehler und unterbreiten identische Korrekturvorschläge.

Außerdem lässt sich in dieser Dialogsequenz beobachten, dass Claude die Sinnhaftigkeit der Antwort der Testperson hinterfragt. Claude weist darauf hin, dass es in einem Geschäft für Brautmoden wahrscheinlich keine Anzüge zu kaufen geben wird. Im Gegensatz dazu problematisiert ChatGPT diese mögliche inhaltliche Diskrepanz nicht, das Programm beschränkt sich auf die im Prompt Feedback-Kriterien, ohne die inhaltliche Logik der Antwort zu hinterfragen.

Tab. 5: Vierte Fragen und Antworten

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Nächste Frage We have dresses in many colors, including red. Would you like to try on some dresses?</p>	<p>Nächste Frage: I see. We have some beautiful red dresses, but they are not traditional for weddings. Would you like to try a white dress with red details?</p>
<p>Antwort der Testperson Thank you, i would like to try some dresses. Please show me the red dresses first.</p>	<p>Antwort der Testperson Let's have a look on this dresses second. Please, show me first the red pieces.</p>
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Es gibt einen kleinen Rechtschreibfehler.</p> <p>Ist der Ausdruck gut und verständlich? Ja, der Ausdruck ist gut und verständlich.</p> <p>Feedback Deine Antwort ist sehr gut verständlich und die Vokabeln sind richtig. Es gibt nur einen kleinen Rechtschreibfehler.</p> <p>Fehlerhinweis Der Satz "Thank you, i would like to try some dresses." enthält einen Rechtschreibfehler: "i" sollte großgeschrieben werden.</p> <p>Korrektur Es sollte heißen: "Thank you, I would like to try some dresses."</p> <p>Perfekte Antwort auf A2-Niveau "Thank you, I would like to try some dresses. Please show me the red dresses first."</p>	<p>Analyse:</p> <ul style="list-style-type: none"> • Vokabeln: Größtenteils passend, aber einige Ungenauigkeiten. • Grammatik: Einige Fehler, z.B. bei Pluralformen und Präpositionen. • Ausdruck: Verständlich, aber mit Verbesserungspotenzial. <p>Feedback: Deine Antwort zeigt, dass du deine Wünsche ausdrücken kannst, aber es gibt einige Punkte zu verbessern:</p> <ol style="list-style-type: none"> 1. "on this dresses" sollte "at these dresses" heißen. "This" ist Singular, "these" ist der korrekte Plural. 2. "second" am Satzende ist ungewöhnlich platziert. Es wäre besser am Anfang des Satzes. 3. Der erste Satz könnte klarer formuliert werden. 4. "pieces" ist nicht das beste Wort für Kleider. "dresses" wäre passender. <p>Korrekte Version: "Let's look at these dresses later. First, please show me the red dresses."</p> <p>Eine perfekte Antwort auf A2-Niveau könnte so aussehen: "Okay. Can we look at the red dresses first, please? We can see the white dresses later."</p>

Beobachtungen und Feststellungen

Die vorhergehenden Feststellungen bestätigen sich erneut bei den Dialogsequenzen der vierten und fünften Frage, die deshalb hier zusammengefasst behandelt werden: Beide Chatbots erkennen Fehler in den Antworten und unterbreiten richtige Korrekturvorschläge. Erneut ist der Chatbot Claude in seinem Feedback variantenreicher als ChatGPT.

Tab. 6: Fünfte Fragen und Antworten

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Nächste Frage Sure, here are some red dresses. Do you prefer a specific material, like silk or lace?</p>	<p>Nächste Frage: Certainly! Here are our red dresses. This one has a beautiful lace pattern. Would you like to try it on?</p>
<p>Antwort der Testperson Oh, I don't know what ist the difference between silk or lace. Please, tell me what the difference is.</p>	<p>Antwort der Testperson I don't like lace. But the the cut is very well. It looks good. Yes, I would like to try it on.</p>
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind größtenteils passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Es gibt einige kleine Fehler.</p> <p>Ist der Ausdruck gut und verständlich? Der Ausdruck ist größtenteils verständlich, aber die Fehler könnten verwirrend sein.</p> <p>Feedback Deine Antwort ist größtenteils verständlich und die Vokabeln sind gut. Es gibt jedoch einige kleine Fehler in der Grammatik und Rechtschreibung.</p> <p>Fehlerhinweis 1. Der Satz "I don't know what ist the difference between silk or lace" enthält einen Fehler im Wort "ist"; es sollte "is" heißen. 2. Anstatt "between silk or lace" sollte es "between silk and lace" heißen.</p> <p>Korrektur Es sollte heißen: "Oh, I don't know what the difference is between silk and lace. Please, tell me what the difference is."</p> <p>Perfekte Antwort auf A2-Niveau "Oh, I don't know what the difference is between silk and lace. Please, tell me what the difference is."</p>	<p>Analyse:</p> <ul style="list-style-type: none"> • Vokabeln: Größtenteils passend, aber einige Ungenauigkeiten. • Grammatik: Kleinere Fehler, z.B. Wiederholung von "the". • Ausdruck: Verständlich, aber mit Verbesserungspotenzial. <p>Feedback: Deine Antwort zeigt, dass du deine Meinung ausdrücken kannst, aber es gibt einige Punkte zu verbessern:</p> <ol style="list-style-type: none"> 1. "the the cut" ist eine Wiederholung. Einmal "the" genügt. 2. "very well" ist nicht die beste Beschreibung für einen Schnitt. "very nice" oder "very good" wäre passender. 3. Der Ausdruck "It looks good" ist korrekt, könnte aber präziser sein. 4. Die Sätze sind etwas abgehackt. Sie könnten besser verbunden werden. <p>Korrekte Version: "I don't like lace, but the cut is very nice. The dress looks good. Yes, I would like to try it on."</p> <p>Eine perfekte Antwort auf A2-Niveau könnte so aussehen: "I don't like lace, but I like the cut of the dress. It looks beautiful. Yes, I want to try it on, please."</p>

Tab. 7: Sechste Frage und Antwort

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Erklärung und nächste Frage Silk is a smooth, soft fabric that feels very luxurious, while lace is a delicate, openwork fabric that is often used for decorative purposes. Do you have a budget in mind for your dress?</p>	
<p>Antwort der Testperson Okay, I prefer a dress made of silk. Money doesn't matter. I'm a very rich man. Please show me the most expansive pieces first.</p>	
<p>Analysiere die Antwort</p> <p>Wurden die passenden Vokabeln genutzt? Ja, die Vokabeln sind größtenteils passend.</p> <p>Sind Grammatik und Rechtschreibung korrekt? Es gibt einige kleine Fehler.</p> <p>Ist der Ausdruck gut und verständlich? Der Ausdruck ist größtenteils verständlich, aber die Fehler könnten verwirrend sein.</p> <p>Feedback Deine Antwort ist größtenteils verständlich und die Vokabeln sind gut. Es gibt jedoch einige kleine Fehler in der Grammatik und Rechtschreibung.</p> <p>Fehlerhinweis 1. Der Satz "Please show me the most expansive pieces first" verwendet "expansive" (weitreichend) statt "expensive" (teuer). 2. Der Satz "Money doesn't matter" enthält ein typographisches Apostroph; es sollte ein gerades Apostroph sein („ ' “).</p> <p>Korrektur Es sollte heißen: "Okay, I prefer a dress made of silk. Money doesn't matter. I'm a very rich man. Please show me the most expensive pieces first."</p> <p>Perfekte Antwort auf A2-Niveau "Okay, I prefer a dress made of silk. Money doesn't matter. I'm a very rich man. Please show me the most expensive pieces first."</p>	

Beobachtungen und Feststellungen

Eine interessante Beobachtung ist, dass ChatGPT – aus unbekanntem Grund – noch eine sechste Frage stellt, während für Claude der Dialog bereits beendet ist. Dieses Verhalten könnte darauf hinweisen, dass die beiden Chatbots die Bedeutung der Eröffnungsfrage unterschiedlich interpretieren und daher zu einer abweichenden Zählweise gelangen.

Inhaltlich gibt es den vorangegangenen Feststellungen nichts hinzuzufügen.

Tab. 8: Zusammenfassendes Resümee der Chatbots

ChatGPT HAWKi-API	Claude 3.5 Sonnet
<p>Zusammenfassendes Resümee Deine Antworten waren insgesamt gut verständlich und die Vokabeln waren größtenteils passend. Es gab einige kleine Grammatik- und Rechtschreibfehler, die du jedoch mit etwas Übung leicht verbessern kannst. Achte besonders auf die richtige Verwendung von Apostrophen und die Unterscheidung zwischen ähnlich klingenden Wörtern (z.B. "expensive" vs. "expansive"). Dein Ausdruck war gut und hat gezeigt, dass du dich klar und verständlich ausdrücken kannst.</p> <p>Ich hoffe, es hat Dir Spaß gemacht und Du konntest etwas lernen.</p>	<p>Zusammenfassendes Resümee: Insgesamt hast du dich gut ausgedrückt und konntest deine Wünsche und Meinungen verständlich kommunizieren. Deine Antworten waren meist relevant und zeigten ein gutes Verständnis des Kontexts.</p> <p>Stärken:</p> <ul style="list-style-type: none"> • Du verwendest relevante Vokabeln zum Thema Kleidung und Einkaufen. • Du kannst grundlegende Sätze bilden, um deine Präferenzen auszudrücken. • Du verstehst die Fragen und reagierst angemessen darauf. <p>Verbesserungspotenzial:</p> <ul style="list-style-type: none"> • Achte auf korrekte Pluralformen und Artikelverwendung. • Übe die Verwendung von Präpositionen, besonders bei Ortsangaben. • Versuche, deine Sätze etwas flüssiger zu verbinden, um den Ausdruck zu verbessern. • Erweitere deinen Wortschatz für präzisere Beschreibungen von Kleidung und Stilen. <p>Um dich zu verbessern, solltest du:</p> <ol style="list-style-type: none"> 1. Regelmäßig Vokabeln zum Thema Mode und Einkaufen üben. 2. Die korrekte Verwendung von Artikeln und Pluralformen trainieren. 3. Kurze Dialoge lesen und schreiben, um deine Ausdrucksfähigkeit zu verbessern. 4. Auf die korrekte Verwendung von Präpositionen achten. <p>Mit weiterer Übung wirst du bald noch sicherer in solchen Gesprächen werden.</p> <p>Ich hoffe, es hat Dir Spaß gemacht und Du konntest etwas lernen.</p>

Beobachtungen und Feststellungen

Das „Zusammenfassende Resümee“ schließt die Dialoge in den Chatbots ab. Es kann festgestellt werden, dass die Bewertungen der Leistung der Testperson durch die Chatbots weitgehend übereinstimmt. Beide Systeme benennen Stärken und Schwächen und bieten Anknüpfungspunkte für gezieltes Sprachtraining, um die Kompetenzen zu verbessern. Dabei existieren in den Nuancen der Rückmeldungen durchaus Differenzen.

Ein bemerkenswerter Punkt ist jedoch, dass Claudes Feedback umfangreicher und besser gegliedert scheint, wodurch es professioneller wirkt. Claude bietet detailliertere Rückmeldungen, die sowohl das Verbesserungspotenzial als auch vier konkrete Handlungsempfehlungen enthalten, mit denen diese Potenziale aktualisiert werden könnten.

5 Diskussion und Fazit

Die vorliegende Untersuchung weist einige methodische Limitationen auf, die bei der Interpretation der Ergebnisse zu berücksichtigen sind. Die Studie beschränkte sich für die Validierung der beiden Hypothesen lediglich auf zwei Chatbots, was die Generalisierbarkeit der Resultate einschränkt. Es ist davon auszugehen, dass eine breitere Palette an KI-Systemen sowie verschiedene Versionen derselben zu differenzierteren Erkenntnissen führen könnten. Insofern kann der vorliegende Bericht nur ein erster Impuls sein, die Prüfung der Qualität und der Relevanz von Antworten verschiedener Chatbots und die darauf fußende Bewertung ihrer Eignung für den Einsatz in der Lehre systematisch fortzuführen.

Die Durchführung der Tests erfolgte zudem ausschließlich in englischer Sprache, in der die Chatbots mutmaßlich besonders versiert sind. Zudem wurde mit dem Sprachniveau A2 ein niedriger Schwierigkeitsgrad gewählt. Eine Ausweitung der Untersuchung auf weitere Sprachen sowie höhere Kompetenzstufen wäre wünschenswert. Hierbei könnte die KI-interessierte Community durch kollaborative Beiträge und Rückmeldungen einen wertvollen Beitrag leisten.

Trotz dieser Einschränkungen lassen sich zwei zentrale Erkenntnisse ableiten: Erstens erwies sich der entwickelte Prompt – unter den genannten Vorbehalten – als chatbot-übergreifend funktional und zufriedenstellend. Der implementierte Korrekturmechanismus operierte, von einer singulären Ausnahme abgesehen, zuverlässig. Die von den KI-Systemen generierten Rückmeldungen können als adäquat bewertet werden.

Zweitens sind nach Meinung des Autorenkollektivs beide initialen Hypothesen bestätigt. Die kleine Untersuchung zeigt, dass eine vergleichende Analyse desselben Prompts in verschiedenen generativen Chatbots gewinnbringend sein und ChatGPT, ungeachtet seiner Prominenz, mit durchaus gleichwertige Konkurrenten im Wettbewerb steht.

Diese Schlussfolgerung stützt sich auf mehrere Beobachtungen: Die Rückmeldungen der Chatbots zeichneten sich durch Diversität und variierenden Umfang aus, wenngleich in der Diagnose und Korrektur der Fehler eine softwareübergreifende Kongruenz festzustellen war. ChatGPT bearbeitete den Prompt nur dann vollständig, wenn in der Antwort der Testperson Fehler identifiziert wurden. Bei fehlerfreien Antworten wurde die Anweisung, eine mustergültige Antwort auf dem geforderten Sprachniveau zu präsentieren, nicht befolgt. Claude hingegen setzte die Vorgaben des Prompts konsequent um, indem es auch bei fehlerfreien Antworten der Testperson eine exemplarische Antwort auf dem Sprachniveau A2 bereitstellte. Diese stringenteren Umsetzung der Anforderungen durch Claude kann als Vorteil in der Evaluation der Chatbot-Leistung gewertet werden. Für Anwender von ChatGPT empfiehlt sich eine Überarbeitung und Optimierung des Prompts.

Darüber hinaus bot Claude eine zusätzliche Ebene der inhaltlichen Überprüfung und demonstrierte die Fähigkeit, die Relevanz und Plausibilität der Antworten kritisch zu hinterfragen. Im Gegensatz dazu thematisierte ChatGPT etwaige inhaltliche Diskrepanzen nicht, was auf eine stärkere Fokussierung auf die formale Korrektheit der Sprache hindeutet, ohne die inhaltliche Kohärenz der Antwort zu problematisieren. Je nach Lehr- und Lernszenario könnte dies zu einer präferenziellen Nutzung des einen oder anderen KI-Systems führen. Es wurde konstatiert, dass Claude nicht durchgängig fehlerfrei korrigierte. In einem spezifischen Fall, in dem das Wort „fully“ einer Korrektur bedurfte, wäre „curvy“ dem vorgeschlagenen „full“ vorzuziehen gewesen. Dies indiziert, dass Claude zwar detaillierte Vorschläge unterbreitet, gelegentlich jedoch Optimierungspotenzial hinsichtlich der Präzision der Korrekturen besteht.

Resümierend lässt sich konstatieren, dass beide evaluierten Chatbots effektive Instrumente zur Fehlerdetektion und Sprachkorrektur darstellen, dabei jedoch divergierende Ansätze in der Rückmeldung verfolgen. Claude offeriert ein breiteres Spektrum an Verbesserungsoptionen, während ChatGPT sich auf die präzise Korrektur der gegebenen Antwort konzentriert. Diese Differenzen können, abhängig von den spezifischen Anforderungen und Zielsetzungen der Nutzer, als vorteilhafte Charakteristika betrachtet werden.

In einem allgemeineren Kontext lässt sich festhalten, dass generative Chatbots im Sprachunterricht gewinnbringend eingesetzt werden können. Mit dem „Dialog-Trainer“-Prompts wurde ein Beispiel präsentiert, das sich insbesondere für selbstgesteuerte Lernsequenzen zum Training von Vokabeln und grammatischen Strukturen eignet. Es zeigt auch, dass mit KI-Systemen freie und abwechslungsreiche Aufgabenstellungen zu realisieren sind, die trotz Abwesenheit einer Lehrperson unmittelbare Korrektur- und Feedbackfunktionen bieten.

Daraus könnten verschiedene Vorteile resultieren, etwa eine Reduktion des Vorbereitungs- und Nachbearbeitungsaufwands für Lehrende. Lernende profitieren hingegen von der zeitlichen und örtlichen Flexibilität bei der Durchführung der Übungen. Auch sollte die Tatsache nicht außer Acht gelassen werden, dass die Dialoge beliebig oft wiederholt werden können, wobei die KI für jeden neuen Trainingsdurchgang – selbst bei identischer Themenstellung – neue Dialogsituationen kreieren würde. Natürlich könnten Lernende durch gezielte Modifikationen des Prompts das Übungs- und Trainingsfeld thematisch variieren.

Die Autoren würden eine Erprobung des Prompts durch die Leserschaft sehr begrüßen und sind an Rückmeldungen zu Erfahrungen und Einschätzungen interessiert. Konstruktive Vorschläge zur Weiterentwicklung und Optimierung des Prompts sind jederzeit willkommen.