

Generative KI zur Lernbegleitung in den Bildungswissenschaften: Implementierung eines LLM-basierten Chatbots im Lehramtsstudium

Hassan Soliman ¹, Milos Kravcik ¹, Alexander Tobias Neumann ², Yue Yin ²,
Norbert Pengel ³, Maike Haag ³ und Heinz-Werner Wollersheim ³

Abstract:

Lehramtsstudierende nehmen an einer Vielzahl von Lernaktivitäten in verschiedenen Fachbereichen teil, u.a. in den Bildungswissenschaften. Sie benötigen angemessene Unterstützung, u.a. in Form zeitnahen Feedbacks. Da ihre Anzahl steigt, ist es ein wichtiges Ziel, ihnen eine skalierbare Unterstützung zu bieten, was auch eine Herausforderung für Forschende und Entwickelnde darstellt. Als eine mögliche Lösung stellen wir einen Chatbot vor, der auf einem Large Language Model (LLM) basiert und durch bestehende Lern- und Informationsmaterialien sowie hochschuldidaktische Prozessbeschreibungen ergänzt wird. In diesem Praxisbeitrag werden Erfahrungen aus der didaktischen Konzeption und technischen Umsetzung sowie erste Ergebnisse zur Qualität von LLM-basierten Chatbot-Antworten auf inhaltliche und organisatorische Fragen eines technologiegestützten bildungswissenschaftlichen Moduls im Lehramtsstudium vorgestellt. Der Bot ist in der Lage, Kursmaterial abzurufen, zu analysieren und umfassende Antworten auf spezifische Fragen zu geben. Vorläufige Ergebnisse zeigen, dass es möglich ist, zwischen verschiedenen Kontexten, in denen die Studierenden arbeiten, zu unterscheiden und ihnen eine schnelle Antwort zu geben, die das relevante Material berücksichtigt.

Keywords: Chatbot, Large Language Model, Skalierbares Mentoring

1 Einleitung

Skalierbares Mentoring ist eine Herausforderung im Bildungswesen. Die Entwicklung entsprechender Technologien, insbesondere der Künstlichen Intelligenz (KI), eröffnet neue Perspektiven in der digitalen Hochschulbildung. Unser Ziel ist es, einen Chatbot zu entwickeln, der Studierende in bildungswissenschaftlichen Modulen unterstützt. Vorhandene Lern- und Informationsmaterialien sowie hochschuldidaktische Prozessbeschreibungen können das Potenzial in Bezug auf Textverständnis und -generierung erhöhen. Hier wurde generative KI eingesetzt, um einen bereits implementierten Chatbot zu verbessern [Ne21].

- 1 DFKI, Berlin, Germany, hassan.soliman@dfki.de,  <https://orcid.org/0009-0003-4574-9074>;
milos.kravcik@dfki.de,  <https://orcid.org/0000-0003-1224-1250>
- 2 RWTH Aachen University, Aachen, Germany,
neumann@dbis.rwth-aachen.de,  <https://orcid.org/0000-0002-9210-5226>;
yue.yin@rwth-aachen.de,  <https://orcid.org/0009-0006-8369-8396>
- 3 Leipzig University, Leipzig, Germany,
norbert.pengel@uni-leipzig.de,  <https://orcid.org/0000-0002-3263-6877>;
maike.haag@uni-leipzig.de,  <https://orcid.org/0009-0001-6161-2022>;
wollersheim@uni-leipzig.de,  <https://orcid.org/0000-0002-4690-5839>

Zunächst wird auf verwandte Arbeiten verwiesen, danach werden die didaktischen Inhalte erläutert. Im Hauptteil wird der technologische Hintergrund vorgestellt, einschließlich des Chatbot-Designs und der Implementierung des LLM-Prototyps. Abschließend werden erste Ergebnisse präsentiert und die nächsten Schritte skizziert.

2 Verwandte Arbeiten

KI kann eine zentrale Rolle im Bildungswesen spielen, da sie in der Lage ist, große Mengen an Daten im Zusammenhang mit dem Lernprozess zu analysieren [Ki23]. Der Aufbau sozialer generativer KI für Bildung erfordert die Entwicklung leistungsfähiger KI-Systeme, die mit Menschen kommunizieren können. Wir müssen überlegen, wie man eine sozial-generative KI für den Bildungsbereich entwickelt und einschränkt [Sh23].

Forschende haben mit generativen KI-Modellen experimentiert, um herauszufinden, welche Aufgaben sie unterstützen können. Beispiele hierfür sind automatisch generierte Online-Kurse zum Sprachenlernen [RP23], die Bewertung der Korrektheit von Schülerantworten und die Erzeugung von Feedback in einem digitalen Lernspiel [Ng23] und die Bewertung von Aufsätzen zur Stärkung der Schreibfähigkeiten von Studierenden [Se23]. Solche Ergebnisse werden zeigen, welche Aufgaben an KI delegiert werden können und welche auch in Zukunft echte menschliche Fähigkeiten erfordern werden.

3 Lernunterstützung in den Bildungswissenschaften

Die Szenarien in unserem Testbed der Universität Leipzig fokussieren sich didaktisch auf die Unterstützung selbstgesteuerten Lernens sowie auf Mentoring und Beratung im Rahmen eines bildungswissenschaftlichen Moduls des Lehramtsstudiums [Pe23]. Zur Unterstützung von Lernprozessen gilt Mentoring als effektives Mittel [ED15]. Ein zentrales Merkmal ist die dyadische Beziehung zwischen Mentor/in und Mentee [Zi09], welche in Lehrveranstaltungen mit vielen Studierenden aufgrund der erforderlichen Ressourcen nicht realisierbar ist. Daher stellt sich die Frage, wie mentorielle Prozesse skaliert werden können, um die Lernunterstützung nach dem Vorbild des Mentorings umzusetzen. Zur Beantwortung dieser Fragestellung wurden lernbegleitende Szenarien in der Mentoring Workbench (MWB), einer digitalen Arbeitsumgebung, die verschiedene Tools bündelt [ZBJ23] für die Unterstützung des Selbststudiums entwickelt. Eine zentrale Komponente der MWB ist der Chatbot, der als virtueller Ansprechpartner rund um die Uhr verfügbar ist und zwei Funktionen erfüllen soll: Experte und Lernbegleiter [DRS18]. Als Lernbegleiter unterstützt er den individuellen Lernprozess der Studierenden. Er gibt Feedback zu den eingereichten Schreibaufgaben und ermutigt die Studierenden, ihren eigenen Lernprozess zu planen, zu überwachen und zu reflektieren. Als Experte beantwortet der Chatbot Fragen zum Inhalt und zur Organisation Moduls im Rahmen des Lehramtsstudiums an. In diesem Beitrag wird letzteres thematisiert.

4 Design und Implementierung

Bis Ende 2022 konzentrierte sich unsere Chatbot-Entwicklung auf Open-Source-Lösungen wie RASA⁴, ein Framework für maschinelles Lernen, das explizit für die Entwicklung konversationeller KI konzipiert wurde. Teile des Frameworks wurden genutzt, um das Social Bot Framework (SBF) zu integrieren. Dies wurde durch die Fähigkeit zur webbasierten, kollaborativen Entwicklung motiviert. Die Plattform bietet eine intuitive Möglichkeit zur Definition von Benutzerintentionen und Bot-Reaktionen durch Drag-and-Drop-Funktionalitäten [NLK19]. RASA NLU (Natural Language Understanding) wurde integriert, um eine solide Grundlage für die Konversation zu schaffen und gleichzeitig die Entwicklungserfahrung zu verbessern.

Das Team aus dem Bereich der Erziehungswissenschaften hat Beispiele für jede identifizierte Absicht und Entität erstellt. Für die NLU-Komponente wurden Modelle wie der Dual Intent and Entity Transformer (DIET) verwendet, eine hochmoderne Architektur zur gleichzeitigen Klassifizierung von Absichten und Entitäten [AAR22]. Dadurch kann der Chatbot relevante Informationen aus den Benutzerinteraktionen analysieren und herausfiltern. Auch der ursprüngliche Chatbot [Ne21] war bereits in die Webschnittstelle MWB integriert. Unser Ansatz zur vorlagengesteuerten natürlichen Sprachgenerierung (NLG) wies trotz dieser Fortschritte einige Einschränkungen und Herausforderungen auf. Insbesondere die begrenzte Flexibilität bei Abfragen, die über die vordefinierten Regeln hinausgehen, sowie das komplexe Konversationsmanagement aufgrund fehlenden Textbewusstseins und die Benutzerfrustration aufgrund einer weniger ansprechenden Benutzererfahrung stellten Herausforderungen dar.

4.1 LLM-basierter Prototyp

Der BiWi-AI-Tutor ist ein Chatbot-Prototyp auf Basis von LLM. Er wurde entwickelt, um eine skalierbare Lernunterstützung zu ermöglichen, indem er Wissen aus Vorlesungsfolien, Seminartexten und Organisationsmaterialien abrufen. Der Chatbot nutzt das Modell gpt-3.5-turbo von OpenAI⁵ und die LangChain⁶-Bibliothek, um nahtlos innerhalb der MWB eine reaktionsschnelle und kontextbewusste dialogische Interaktion zu bieten. Wir haben den Function Calling Agent in der LangChain-Bibliothek verwendet, eine leistungsstarke Funktion, die es großen Sprachmodellen (LLMs) ermöglicht, spezifische Funktionen oder Werkzeuge als Teil ihres Verarbeitungsworkflows aufzurufen. Diese Agenten können die Fähigkeiten von LLMs erweitern, indem sie Aufgaben ausführen, die externe Funktionen oder den Zugriff auf bestimmte Werkzeuge erfordern. Er wählt die am besten geeigneten Materialien für eine gegebene Anfrage aus. Durch die Kombination der generativen Leistung eines LLM mit Retrieval-Systemen findet das System effizient relevante Abschnitte der

4 <https://rasa.com/>

5 <https://platform.openai.com/docs/models/gpt-3-5-turbo>

6 <https://www.langchain.com/>

Lerninhalte, aus denen es Antworten generiert. Dies wird als Retrieval Augmented Generation (RAG) bezeichnet. Die Effektivität wird durch die Einbeziehung von direktem Feedback der Nutzenden, das durch ein Bewertungssystem (Daumen hoch oder Daumen runter) und Kommentare innerhalb des Chat-Interfaces gesammelt wird, erhöht. Die Interaktionsarchitektur des Chatbots ist sorgfältig aufgebaut, um ein benutzerfreundliches Erlebnis zu gewährleisten. Sie folgt einer logischen Abfolge, beginnend mit der Nachricht des Nutzers, dem Verstehen und Verarbeiten der Anfrage und der Auswahl des richtigen Werkzeugs, je nach Art der Frage. Die verwendeten Tools sind textlich beschrieben und dienen als Indizes für die relevantesten Inhalte, sodass der Chatbot relevante und präzise Informationen abrufen kann. Ein zentraler Aspekt der Funktionalität des Chatbots ist die Indizierung von Inhalten durch den innovativen Einsatz von Inhalteinbettungen. Textabschnitte werden mithilfe der Einbettungsmodelle von OpenAI zu Einbettungen verarbeitet und in einer Vektordatenbank gespeichert. Nach Erhalt einer Anfrage generiert der Chatbot eine semantisch reichhaltige Vektordarstellung der Frage. Dieser wird verwendet, um kontextuell relevante Texte aus den Inhaltsindizes durch semantische Ähnlichkeits- und Schlüsselwortabgleichsmaßnahmen abzurufen. Ein hybrider Ensemble-Retriever-Ansatz wird verwendet, indem ein semantischer Vektor-Retriever über eine Vektordatenbank und ein Schlüsselwort-Retriever über den BM25-Algorithmus [RZ09] eingesetzt wird. Wir haben die fünf besten Abschnitte vom Vektor-Retriever mit den fünf besten Abschnitten vom Schlüsselwort-Retriever abgerufen, um insgesamt zehn Textfragmente für die Abfrage zu erhalten, die als Kontext und Wissen für das LLM dienen. Nach dem Abruf des relevanten Kontexts konstruiert der Chatbot eine fundierte Antwort auf die Anfrage des Nutzenden.

4.2 Ergebnisse

Für die Evaluation des LLM-basierten Chatbots wurde ein Datensatz verwendet, der Fragen aus der Lehrveranstaltung, entsprechende richtige Antworten und die vom Chatbot generierten Antworten enthält. Der Datensatz wurde kuratiert, um die Vielfalt der Materialien widerzuspiegeln. Diese umfassen Vorlesungsfolien, Seminartexte und organisatorische Informationen, die in drei verschiedene Typen kategorisiert und verarbeitet wurden. Die Materialien umfassen Vorlesungsfolien, Seminartexte und organisatorische Informationen im PDF-Format. Sie wurden mithilfe des Moduls LlamaParse aus der Bibliothek LlamaIndex⁷ geparkt, gechunked und indiziert. Dabei wurde für jede Kategorie ein eigener Index erstellt. Zusätzlich wurde ein weiterer Index erstellt, der alle Kategorien enthält. Während des Indexierungsprozesses wurden die PDF-Dateien in Chunks von 1024 Token unterteilt, wobei eine Überlappung von 20 Token verwendet wurde. Anschließend wurden diese Chunks mithilfe des 'text-embedding-3-small'-Modells von OpenAI⁸ in Einbettungen umgewandelt und in einer Vektordatenbank gespeichert, um sie abrufen zu können. Der Evaluierungsdatensatz bestand aus Fragen, die von den Lehrenden des bildungswissenschaftlichen Lehramtsmoduls auf Basis vorhandener Materialien entwickelt wurden, um

⁷ <https://www.llamaindex.ai/>

⁸ <https://platform.openai.com/docs/guides/embeddings/embedding-models>

die Fähigkeit des Chatbots, Antworten zu generieren, zu bewerten. Der Datensatz wurde mit dem QA-Evaluierungsprompt aus der LangChain-Bibliothek ausgewertet. Dabei wurde das gpt-4-Modell von OpenAI⁹ verwendet, um die Korrektheit der Antworten des Chatbots zu beurteilen. Der Bewertungsprozess folgte einer strukturierten Eingabeaufforderung, die die sachliche Richtigkeit betonte und stilistische sowie formale Unterschiede zwischen der Ausgabe des Chatbots und der tatsächlichen Antwort außer Acht ließ. Der Bewertungsmechanismus orientierte sich an der folgenden Aufforderung an gpt-4:

”You are a teacher grading a quiz. Grade the student answers based ONLY on their factual accuracy. Ignore differences in punctuation and phrasing between the student answer and true answer. It is OK if the student answer contains more information than the true answer, as long as it does not contain any conflicting statements. Begin!”

Die Leistung wurde für jede Materialkategorie separat bewertet. Es wurden zwei Ergebnisgruppen ermittelt: eine unter der Annahme, dass das richtige Lernmaterial zur Verfügung gestellt wurde, und ein weiterer, der die Fähigkeit des Chatbots bewertet, das richtige Lernmaterial auszuwählen, wenn alle Lernmaterialien für alle Kategorien als ein Index oder Tool gegeben sind. Die Ergebnisse zeigen, dass der Chatbot in der Lage ist, kontextuell relevante und sachlich korrekte Antworten zu generieren, wenn der passende Lernmaterialindex bereitgestellt wird. Sie verdeutlichen (Tab. 2), dass der Chatbot bei der Bereitstellung des entsprechenden Tools für die jeweilige Kategorie in 85-90% der Fälle korrekt antwortet.

Jedoch sinkt die Anzahl der richtigen Antworten, wenn alle Lernmaterialien in einem einzigen Index zusammengefasst werden. Die Wahl des richtigen Lernmaterials, wenn alle Materialien in einem Index gegeben sind stellt eine kritische Herausforderung für zukünftige Erweiterungen dar, um das kontextuelle Verständnis und die Entscheidungsfähigkeit des Chatbots zu verbessern. Diese Bewertung betont das Potenzial des Chatbots als akademisches Hilfsmittel und zeigt gleichzeitig Bereiche auf, in denen weitere Verbesserungen erforderlich sind, insbesondere bei der Analyse und Auswahl des richtigen und relevanten Lernmaterials, wenn alle Materialien in einem Index gegeben sind.

Tab. 1: Bewertungsergebnisse mit dem Anteil der richtigen Antworten bei ausschließlicher Bereitstellung des passenden Materials und bei Bereitstellung aller Materialien

Kategorie	Anzahl Fragen	Passendes Material	Alle Materialien
Seminar	20	17 (85%)	15 (75%)
Vorlesung	20	17 (85%)	17 (85%)
Organisatorisches	20	18 (90%)	17 (85%)

Tab. 2 zeigt jeweils ein Beispiel aus Vorlesung und Seminar. Beim Beispiel der Vorlesung wurde die Antwort des Chatbots im Vergleich mit der vorgegebenen Antwort als richtig bewertet, beim Beispiel des Seminars als falsch.

9 <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>

Tab. 2: Beispiele für Chatbot-Antworten

Beispiel	Details
Richtig	<p>Frage: August Hermann Francke forderte, Kinder mit aller Freundlichkeit, Milde und unermüdlicher Geduld zur Liebe zur Wahrheit, Gehorsam und Fleiß anzuhalten. Wie stand Francke zu körperlichen Strafen?</p> <p>Richtige Antwort: Verstocktheit und Trotz dürfen mit körperlichen Strafen gebrochen werden.</p> <p>Chatbot-Antwort: August Hermann Francke war der Ansicht, dass Kinder mit aller Freundlichkeit, Milde und unermüdlicher Geduld zur Liebe zur Wahrheit, Gehorsam und Fleiß angehalten werden sollten. Er glaubte, dass nur auf diese Weise das kindliche Herz dem Erzieher in Liebe öffnet. Francke war auch der Meinung, dass Verstocktheit und Trotz mit körperlichen Strafen gebrochen werden dürfen.</p>
Falsch	<p>Frage: Wolfgang Klafki versucht den Widerstreit formaler und materialer Bildungstheorien zu überwinden. Welchen Namen gibt er diesem Bildungskonzept?</p> <p>Richtige Antwort: Kategoriale Bildung</p> <p>Chatbot-Antwort: Wolfgang Klafki gibt diesem Bildungskonzept den Namen "Allgemeinbildung".</p>

5 Zusammenfassung

Zur Unterstützung der Studierenden haben wir im Rahmen der digitalen Arbeitsumgebung Mentoring Workbench (MWB) Chatbots entwickelt und erprobt. Wir verwendeten das Social Bot Framework und RASA, um regelbasierte Antworten zu generieren. Antworten auf häufig gestellte Fragen oder Kursinhalte waren jedoch fest vorgegeben. Durch den Einsatz von LLM bei der Chatbot-Entwicklung ist es nun möglich, von einer flexibleren und adaptiveren Antwortgenerierung zu profitieren, die eine tiefere Kontextualisierung ermöglicht. Dies führt zu einer präziseren und personalisierteren Nutzungserfahrung und ermöglicht es den Studierenden, komplexere Fragen zu stellen. Die nächsten Schritte umfassen eine umfassende Evaluierung mit den Nutzenden, auch unter Einbeziehung menschlicher Begutachtung des Chatbot-Outputs und eine entsprechende kontinuierliche Weiterentwicklung. Konzeptionell wird es in den nächsten Schritten u.a. darum gehen, LLM-gestützte Interaktionen des Chatbots zu ermöglichen, die sich an den Prinzipien der Online-Beratung orientieren [MP24].

Danksagung. Die Forschung, die zu diesen Ergebnissen geführt hat, wurde vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projekts "Personalisierte Kompetenzentwicklung und hybrides KI-Mentoring" (tech4compKI) (Förderkennzeichen 16DHB2206, 16DHB2208, 16DHB2213) gefördert.

Literaturverzeichnis

- [AAR22] Arevalillo-Herraez, M.; Arnau-Gonzalez, P.; Ramzan, N.: On Adapting the DIET Architecture and the Rasa Conversational Toolkit for the Sentiment Analysis Task. *IEEE Access* 10, S. 107477–107487, 2022, ISSN: 2169-3536.
- [DRS18] Dyrna, J.; Riedel, J.; Schulze-Achatz, S.: Wann ist Lernen mit digitalen Medien (wirklich) selbstgesteuert? Ansätze zur Ermöglichung und Förderung von Selbststeuerung in technologieunterstützten Lernprozessen. 2018.
- [ED15] Eby, L. T.; Dolan, E. L.: Mentoring in postsecondary education and organizational settings. 2015.
- [Ki23] Kizilcec, R.F.: To advance AI use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education*, S. 1–8, 2023.
- [MP24] Martin, A.; Pengel, N.: Beratung via Chatbot? Möglichkeiten und Anforderungen beim Einsatz Generativer KI in einem bildungswissenschaftlichen Lehramts-Modul. *e-beratungsjournal.net*, Zeitschrift für Onlineberatung und computervermittelte Kommunikation, 2024.
- [Ne21] Neumann, A. T.; Arndt, T.; Köbis, L.; Meissner, R.; Martin, A.; de Lange, P.; Pengel, N.; Klamma, R.; Wollersheim, H.-W.: Chatbots as a Tool to Scale Mentoring Processes: Individually Supporting Self-Study in Higher Education. *Frontiers in Artificial Intelligence* 4, S. 64–71, 2021.
- [Ng23] Nguyen, H. A.; Stec, H.; Hou, X.; Di, S.; McLaren, B. M.: Evaluating chatgpt’s decimal skills and feedback generation in a digital learning game. In: *European Conference on Technology Enhanced Learning*. Springer, S. 278–293, 2023.
- [NLK19] Neumann, A. T.; de Lange, P.; Klamma, R.: Collaborative Creation and Training of Social Bots in Learning Communities. In: *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*. S. 11–19, 2019.
- [Pe23] Pengel, N.; Soliman, H.; Neumann, A.; Kravcik, M.; Yin, Y.; Bez, M.; Stechert, L.; Haag, M.; Martin, A.; Köbis, L.; J., Z.; Jalilov, O.; Meissner, R.; Pögel, A.; Köhler, T.; Wollersheim, H.-W.: Generative KI in der Hochschulbildung: Konzeption und Implementierung eines LLM-gestützten Chatbots. In: *Tagungsband der Jahrestagung der GMW und CampusSource*. 2023.
- [RP23] Rüdian, S.; Pinkwart, N.: Auto-generated language learning online courses using generative AI models like ChatGPT. 2023.
- [RZ09] Robertson, S.; Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3 (4), S. 333–389, 2009.
- [Se23] Seßler, K.; Xiang, T.; Bogenrieder, L.; Kasneci, E.: PEER: Empowering Writing with Large Language Models. In: *European Conference on Technology Enhanced Learning*. Springer, S. 755–761, 2023.
- [Sh23] Sharples, M.: Towards social generative AI for education: theory, practices and ethics. *Learning: Research and Practice* 9 (2), S. 159–167, 2023.
- [ZBJ23] Zawidzki, J.; Bez, M.; Jalilov, O.: Needs and requirements of teachers, learners and administrative staff for AI based mentoring tools in higher education. In: *INTED2023 Proceedings*. IATED, S. 4674–4679, 2023.
- [Zi09] Ziegler, A.: Mentoring: konzeptuelle grundlagen und wirksamkeitsanalyse. *Mentoring: Theoretische hintergründe, empirische befunde und praktische anwendungen*, S. 7–29, 2009.