Systematische Evaluation prompt-basierter, durch KI generierte Multiple-Choice-Fragen in der Medizinischen Biochemie[1]

Masoud Darabi[2] und Shabnam Fayezi[3]

Künstliche Intelligenz (KI)-Sprachmodelle eröffnen neue Möglichkeiten in der medizinischen Ausbildung und können u. a. zur schnellen Erstellung von Multiple-Choice-Fragen (MCQs) dienen. Diese Studie bewertete prompt-basierte, KI-generierte MCQs im Vergleich zu Dozierenden-generierten Fragen in der medizinischen Biochemie hinsichtlich Qualität, kognitivem Niveau und klinischer Relevanz. Hierzu erzeugten drei KI-Plattformen jeweils 20 MCQs zu Stoffwechsel und Krankheitsprozessen. Ergänzt um 20 von Dozierenden erstellte Fragen wurden alle Fragen von fünf Expert:innen anhand eines strukturierten Rasters bewertet. ChatGPT-4 schnitt insgesamt am besten ab; Gemini 2.5 Pro lieferte klinisch relevantere Fragen; Copilot bot die klarsten Vorlagen. Häufige Schwächen waren mehrdeutige Fragestellungen und redundante Formulierungen. Dieser Ansatz ermöglicht effiziente Fragengenerierung, wobei Stärken und Grenzen der jeweiligen Plattform die Auswahl je nach Bildungsziel beeinflussen.

Systematic Evaluation of prompt-based AI-generated multiple-choice questions in medical biochemistry

Artificial intelligence (AI) language models offer new opportunities in medical education, including rapid generation of multiple-choice questions (MCQs). This study evaluated prompt-based AI-generated MCQs in medical biochemistry against instructor-designed items for quality, cognitive level, and clinical relevance. Three AI platforms each generated 20 questions on metabolism and disease processes. Five experts assessed MCQs using a structured rubric. ChatGPT-4 performed best overall; Gemini 2.5 Pro produced more clinically relevant MCQs; Copilot provided the clearest templates. Common weaknesses included ambiguous questions and redundant wording. This approach enables efficient MCQ generation, and platform-specific strengths and limitations influence selection depending on the educational objective.

1 Basiert auf einem Impulsbeitrag im Rahmen der Tagung.

2 ORCID: 0000–0001-6380-272X

3 ORCID: 0000–0001-6954-5138

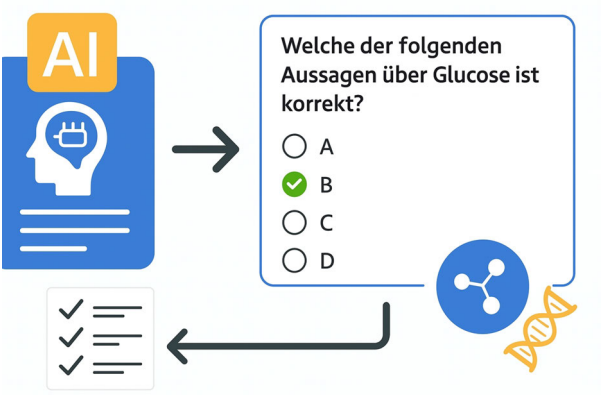
Einleitung

Künstliche Intelligenz (KI), insbesondere große Sprachmodelle (Large Language Models, LLMs), gewinnt in der medizinischen Ausbildung rasch an Bedeutung und verändert vor allem Lehr- und Prüfungsprozesse sowie die Erstellung von Lernmaterialien. Zu den meistgenannten Potenzialen zählt die Möglichkeit, innerhalb kürzester Zeit eine große Anzahl vielfältiger Prüfungsfragen zu generieren und damit die Arbeitsbelastung von Lehrenden zu reduzieren (Abd-alrazaq et al., 2023).

Aktuelle Studien untersuchen den Einsatz von LLMs zur automatisierten Erstellung von Multiple-Choice-Fragen (MCQs), häufig unter Verwendung von Retrieval Augmented Generation (Grévisse et al., 2024; Hang et al., 2024; Kunuku & Dehbozorgi, 2025). Dabei werden, um Qualität und Verständlichkeit zu bewerten, Fragen aus Lehrtexten und Kursmaterialien abgeleitet und anschließend mit manuell erstellten Fragen verglichen. Im Gegensatz dazu basiert die Prompt-basierte, auch als »Closed-Book« bezeichnete, Generierung ausschließlich auf den in den Modellparametern aus den Trainingsdaten abgeleiteten Informationen, ohne externe Quellen abzurufen. Beide Ansätze zeigen Potenzial, unterscheiden sich jedoch hinsichtlich Nachvollziehbarkeit und curricularer Anwendbarkeit. Moderne LLMs sind auf umfangreichen Datensätzen trainiert, was wiederum eine hohe Leistungsfähigkeit in der Sprachverarbeitung sowie die Bearbeitung fachspezifischer Aufgaben, etwa im medizinischen Bereich, ermöglicht. Daher gewinnen prompt-basierte Verfahren, insbesondere aufgrund ihrer Geschwindigkeit, Anpassungsfähigkeit und einfachen Integration in digitale Lernumgebungen, zunehmend an Bedeutung für die akademische MCQ-Erstellung (Abb. 1).

Frühere Arbeiten zeigen, dass digitale Lehrmethoden – insbesondere im Bereich der medizinischen Biochemie – Kreativität fördern und die Motivation der Studierenden steigern können (M. Darabi et al., 2013; M. Darabi & Darabi, 2011). Aufgrund ihrer konzeptionellen Komplexität und der engen Verknüpfung mit der klinischen Praxis stellt die medizinische Biochemie einen besonders anspruchsvollen Prüfstein für die didaktische Validität KI-generierter MCQs dar. Die vorliegende Studie geht daher der Frage nach, inwieweit KI-generierte MCQs den didaktischen Anforderungen der medizinisch-biochemischen Ausbildung gerecht werden können.

Abbildung 1: Workflow der promptbasierten (»Closed-Book«) MCQ-Generierung mit einem LLM: Prompt → MCQ-Entwurf → Expert:innenbewertung vor dem Einsatz in der Lehre.



Methodik

Diese Evaluationsstudie verwendete ein vergleichendes Design (siehe Abb. 2), um die Leistungsfähigkeit von drei webbasierten KI-Plattformen zu bewerten: Gemini 2.5 Pro, Microsoft Copilot (GPT-4 Turbo) und ChatGPT-4. Mithilfe eines standardisierten Prompts generierte jede Plattform 20 MCQs zu drei zentralen Themenbereichen der medizinischen Biochemie: (i) biochemische Signalwege, (ii) Stoffwechselregulation und (iii) krankheitsbezogene Mechanismen. Die Fragen wurden so konzipiert, dass sie, gemäß der Bloom'schen Taxonomie, unterschiedliche kognitive Niveaus abbilden – von Reproduktion und Verständnis bis hin zu Anwendung und Analyse. Anschließend wurden die Fragen von Expert:innen entsprechend diesen vier Niveaustufen kategorisiert.

> **Prompt (verbatim):** Generate multiple-choice questions (MCQs) in medical biochemistry for undergraduate medical students. Create: 5 MCQs on biochemical pathways; 8 MCQs on metabolic regulation; 7 MCQs on disease-related mechanisms. Instructions: Each question must have one correct answer and three plausible distractors (a, b, c, or d). Ensure a mix of cognitive levels: ~30 % recall-based; ~30 % understanding; ~20 % application; ~20 % analysis.

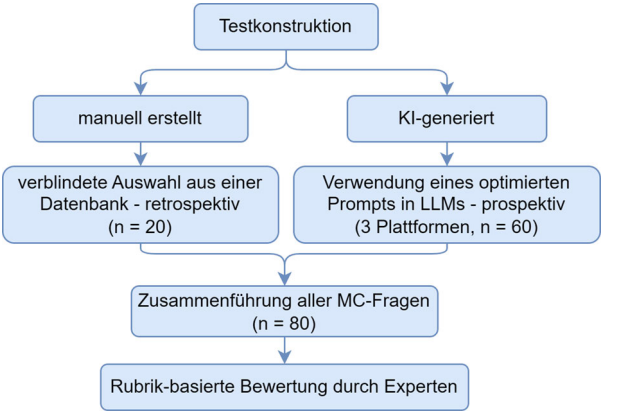
Insgesamt wurden demnach 60 KI-generierte MCQs mit 20 von Dozent:innen erstellten Fragen verglichen. Die menschlich erstellten Fragen wurden zufällig aus einem validierten institutionellen Fragenpool ausgewählt, um inhaltliche Relevanz und Diversität sicherzustellen. Fünf unabhängige Expert:innen aus der Biochemie

und medizinischen Didaktik führten, ohne vorherige Kenntnis über die Herkunft der jeweiligen Fragen, eine verblindete Bewertung durch.

Ein siebenkriterielles Bewertungsraster wurde verwendet: (1) Korrektheit der richtigen Antwort, (2) Plausibilität der Distraktoren, (3) Verständlichkeit der Fragestellung, (4) Item-Schwierigkeit (zielgruppenbezogen; unabhängig von Bloom), (5) sprachliche Qualität, (6) Lernzielrelevanz (Modulhandbuch) und (7) Vermeidung von Hinweisen (keine Cues im Fragetext/Antwortoptionen).

Gruppenunterschiede wurden mit dem Kruskal–Wallis-Test geprüft; Paarvergleiche erfolgten mittels Mann–Whitney-U-Test (Holm-korrigiert).

Abbildung 2: Schematische Übersicht des Studiendesigns. Bewertet wurden 80 MCQs (je 20 KI-generierte aus drei webbasierten LLM-Plattformen sowie 20 von Dozent:innen erstellte). Fünf Fachexpert:innen beurteilten alle Fragen anhand eines siebenkriteriellen Rasters; zudem wurden sie einer von vier Bloom-Stufen zugeordnet.



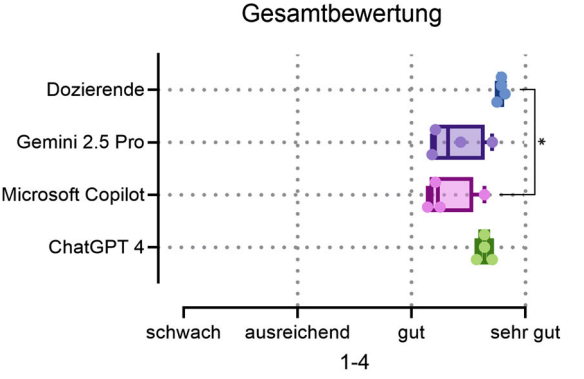
Ergebnisse

Insgesamt zeigten alle KI-Plattformen eine zufriedenstellende Leistung. Die von Dozent:innen erstellten MCQs erhielten, innerhalb der verblindeten Auswertung, die beste Gesamtbewertung, wobei die Punktverteilung deutlich in Richtung der Kategorie »Sehr gut« tendierte. Unter den KI-Plattformen kam ChatGPT-4 der Qualität menschlich generierter MCQs am nächsten.

Gemini 2.5 Pro und Microsoft Copilot erzielten ebenfalls Ergebnisse im Bewertungsbereich »Gut«, zeigten jedoch eine größere Variabilität über die Bewertungskriterien hinweg, was auf eine weniger konsistente Output-Qualität hinweist. Zwi-

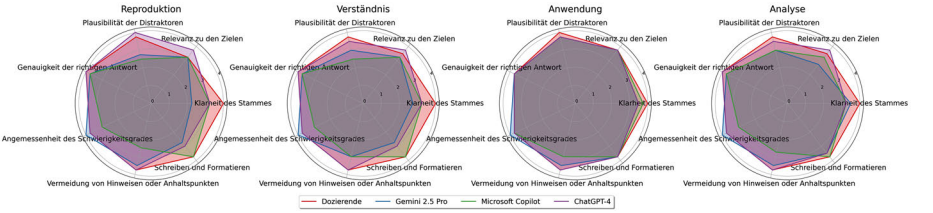
schen den menschlich erstellten Fragen und den von Gemini 2.5 Pro generierten MCQs zeigte sich ein statistisch signifikanter Unterschied ($p < 0{,}05$; siehe Abb. 3).

Abbildung 3: Vergleich der Gesamtbewertung von MCQs, erstellt von menschlichen Autor:innen und KI-Plattformen. Die Bewertungen erfolgten durch Expert:innen anhand einer vierstufigen ordinalen Skala (1 = schwach, 2 = ausreichend, 3 = gut, 4 = sehr gut). Boxplots zeigen Median und Interquartilsbereich; Datenpunkte entsprechen den Mittelwerten der fünf Expert:innenbewertungen pro kognitivem Niveau.



Die Analyse des kognitiven Anspruchsniveaus sowie der sieben Bewertungskriterien zeigte, dass ChatGPT insgesamt die beste Leistung erbrachte – insbesondere hinsichtlich der Korrektheit der Antworten und der Plausibilität der Distraktoren. Gemini 2.5 Pro erzielte schlechtere Bewertungen in Bezug auf die Verständlichkeit der Fragestellung und die Plausibilität der Distraktoren, während Microsoft Copilot die größten Schwierigkeiten bei der Generierung plausibler Distraktoren aufwies (siehe Abb. 4).

Abbildung 4: Vergleich der Leistung von drei KI-Plattformen und manuell erstellten Dozierendenfragen über vier Bloom-Stufen. Bewertet wurden sieben Bewertungskriterien: Korrektheit, Distraktor-Plausibilität, Verständlichkeit, (zielgruppenbezogener) Schwierigkeitsgrad, sprachliche Qualität, Lernzielrelevanz und Vermeidung von Hinweisen. AdS = Angemessenheit des Schwierigkeitsgrads; GdrA = Genauigkeit der richtigen Antwort; KdS = Klarheit des Stamms; PdD = Plausibilität der Distraktoren; RzdZ = Relevanz zu den Lernzielen; SuF = Sprache und Formatierung; VvHoA = Vermeidung von Hinweisen oder Anhaltspunkten.



Auf den niedrigeren kognitiven Ebenen wie Reproduktion und Verständnis schnitten sowohl Gemini als auch Copilot schlechter ab als die menschlich erstellten Fragen und ChatGPT. Besonders bei Gemini beeinträchtigten unklare Fragestellungen die Verständlichkeit. Bei Fragen auf der Bloom-Stufe Analyse zeigten Gemini und Copilot weiterhin Schwächen hinsichtlich der Plausibilität der Distraktoren und der Vermeidung unbeabsichtigter Hinweise, während ChatGPT und die Dozierendenfragen hier ausgewogenere Ergebnisse lieferten. Im Bereich der Anwendung war die Leistung über alle KI-Plattformen hinweg relativ konsistent, mit Ergebnissen, die sich der Qualität menschlich erstellter MCQs annäherten. Gemini erzeugte tendenziell klinisch reichhaltigere Szenarien, während die Stärke von Copilot in der Generierung strukturierter Vorlagen lag (siehe Abb. 4).

Diskussion

Die vorliegende Studie zeigt, dass LLM-basierte KI-Plattformen – Gemini 2.5 Pro, Microsoft Copilot und ChatGPT-4 – MCQs von angemessener Qualität im Bereich der medizinischen Biochemie generieren können, wobei sich je Plattform unter dem verwendeten standardisierten Prompt unterschiedliche Stärken im Output zeigten. Gemini 2.5 Pro lieferte klinisch reichhaltige Fragestellungen, Copilot zeigte einen Schwerpunkt auf Fragen zu biochemischen Signalwegen, häufig in vorstrukturierter Form, und ChatGPT-4 integrierte themenübergreifende Inhalte in prägnanter und kohärenter Formulierung. Auf höheren kognitiven Ebenen entsprachen die KI-generierten Fragen häufig der Qualität menschlich erstellter MCQs oder übertrafen diese sogar, was wiederum ein besonderes Potenzial für den Einsatz in der Bewertung der kognitiven Domänen Anwendung und Analyse erken-

nen lässt. Zu den häufigen Schwächen zählten jedoch unklare oder unzureichend spezifizierte Fragestellungen, Distraktoren, die redundant (inhaltlich sehr ähnlich) oder im Schwierigkeitsniveau unpassend waren, sowie gelegentliche Defizite in der klinischen Relevanz. Insgesamt kann prompt-basierte KI eine wertvolle Ergänzung zu traditionellen Fragenpools darstellen, bedarf aber zum aktuellen Zeitpunkt noch der wissenschaftlichen Einordnung und fachlichen Überprüfung von Expert:innen. Die unterschiedlichen Stärken der einzelnen Plattformen verdeutlichen die Relevanz einer kontextbezogenen Auswahl geeigneter KI-Tools in Abhängigkeit vom Bildungsziel und kognitivem Anspruchsniveau. Es sollte darauf hingewiesen werden, dass die Outputqualität nicht nur vom jeweiligen System, sondern auch von der Promptgestaltung und der plattformspezifischen Verarbeitung abhängt; der hier verwendete standardisierte Prompt könnte daher die beobachteten Unterschiede beeinflusst haben. In Zukunft sind weitere Studien erforderlich, um die Promptsensitivität plattformübergreifend zu quantifizieren.

Praktische Implikationen

Die Ergebnisse legen nahe, dass KI die Vielfalt der Aufgabenpools in der medizinischen Ausbildung erweitern kann. Jede Plattform zeigte spezifische Stärken: Gemini erwies sich als besonders effektiv bei der Generierung klinisch reichhaltiger, fallbasierter Fragestellungen; ChatGPT überzeugte durch prägnante und integrative Fragen zu verschiedenen Themenbereichen; und Microsoft Copilot bot strukturierte Vorlagen, die als nützliche Ausgangspunkte für die wissenschaftliche Praxis dienten. Diese Unterschiede unterstreichen die Notwendigkeit von Transparenz, wenn KI zur Prüfungsentwicklung beiträgt: Studierende sollten über den Einsatz informiert werden, und die verwendeten Prompts sowie zentrale Rahmenbedingungen (z.B. Modell/Version, Datum, Einstellungen) sollten dokumentiert werden, um auch zukünftig methodische Nachvollziehbarkeit und reproduzierbare Qualität zu gewährleisten.

Fazit und Ausblick

KI-generierte MCQs können die medizinische Ausbildung sinnvoll unterstützen, sofern sie sorgfältig integriert werden. Prompt-basierte Ansätze zeigen vielversprechendes Potenzial für den direkten Einsatz in Lehre und Prüfung, häufig mit nur geringem Nachbearbeitungsaufwand. Dennoch sind Expert:innen-basierte Überprüfungen entscheidend, um fachliche Bildungsstandards zu wahren und Verzerrungen vorzubeugen.

Zukünftige Forschung sollte untersuchen, wie sich KI-generierte Prüfungsfragen auf Lernergebnisse auswirken, wenn sie in realen Prüfungssituationen eingesetzt werden. Zudem sollten Strategien zur Minimierung von Bias entwickelt und evaluiert werden. Die Kombination komplementärer Stärken verschiedener KI-Plattformen könnte robuste hybride Workflows ermöglichen, die sowohl die Effizienz als auch die Qualität medizinischer Prüfungsformate verbessern.

Literatur

Abd-alrazaq, Alaa/AlSaad, Rawan/Alhuwail, Dari/Ahmed, Ahmed/Healy, Padraig M./Latifi, Syed/Aziz, Sarah/Damseh, Rafat/Alabed Alrazak, Sadam/Sheikh, Javaid (2023): »Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions«, in: JMIR Medical Education, 9, e48291.

Darabi, M./Darabi, M. (2011): Study on electronic designing and submission for Clinical Biochemistry course assignments as a teaching method, in: Tabriz University of Medical Sciences, 4th National Conference of E-Learning in Medical Sciences.

Darabi, M./Yazdi, Z./Darabi, M./Fayezi, S./Askari, M./Sarchami, R. (2013): »Infrastructure and Faculty Member Readiness for E-learning Implementation: The Case of Qazvin University of Medical Sciences«, in: Iranian Journal of Medical Education, 13(9), S. 730–740. https://ijme.mui.ac.ir/article-1-2695-fa.html.

Grévisse, Christian/Pavlou, Maria A. S./Schneider, Jochen G. (2024): »Docimological Quality Analysis of LLM-Generated Multiple Choice Questions in Computer Science and Medicine«, in: SN Computer Science, 5(5), S. 636.

Hang, Ching N./Wei Tan, Chee/Yu, Pei-Duo (2024): »MCQGen: A Large Language Model-Driven MCQ Generator for Personalized Learning«, in: IEEE Access, 12, S. 102261–102273.

Kunuku, Mourya T./Dehbozorgi, Narsin (2025): »Exploring Multimodal Quiz Generation and Evaluation Aligned with Higher-Order Learning Objectives in Bloom's Taxonomy«. in: Alexandra I. Cristea/Erin Walker/Yu Lu/Olga C Santos/Isotani, Seji, Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED, 2590, S. 433–438.