Public Management and Governance Review – 2026
DOI: https://doi.org/10.60733/PMGR.2026.02

From the class room

Wenn künstliche Intelligenz medizinische Prüfungs- und Versorgungslogiken herausfordert

Thomas Gönner[1]

[1] Universitätsklinikum Wien

Abstract

Große Sprachmodelle wie Google Gemini erreichen zunehmend Leistungsniveaus, die klassische Wissensprüfungen in der medizinischen Ausbildung infrage stellen [1], [2], [3]. Während bisherige Studien vor allem internationale oder englischsprachige Prüfungsformate analysierten, fehlten belastbare empirische Befunde für den deutschsprachigen Raum. Dieser Artikel untersucht erstmals systematisch die Leistungsfähigkeit von Google Gemini anhand von 1.252 originalen Prüfungsfragen des Instituts für medizinische und pharmazeutische Prüfungsfragen (IMPP). Die Ergebnisse zeigen eine Gesamt-Trefferquote von 96 Prozent und damit eine deutliche Überschreitung der üblichen Bestehensgrenze [7], [8]. Der Beitrag interpretiert diese Befunde nicht als unmittelbaren Reformdruck für Prüfungsordnungen, sondern als Anlass zur Reflexion bestehender Prüfungs-, Ausbildungs- und Versorgungslogiken. Im Fokus stehen Governance-, Steuerungs- und Gestaltungsfragen für Entscheidungsträger, Gesundheitspolitik und klinischer Praxis. Abschließend werden Handlungsempfehlungen formuliert.

Key words

Künstliche Intelligenz, Medizinische Ausbildung, Staatsexamina, Klinische Entscheidungsunterstützung, Gesundheitsgovernance

Danksagung

Mein Dank gilt ao. Univ.-Prof. Dr. Johannes Steyrer für die Betreuung der zugrunde liegenden Masterarbeit.

Public Management and Governance Review – 2026
DOI: https://doi.org/10.60733/PMGR.2026.02

Ergänzende Information

Die empirische Datenerhebung und Auswertung dieses Beitrags wurde im Zuge der Masterarbeit im EMBA-Programm Health Care Management an der WU Executive Academy durchgeführt. Datengrundlage sind 1.252 IMPP-Prüfungsfragen aus dem Jahr 2024, bezogen über die Lernplattform AMBOSS, sowie der offizielle Lösungsschlüssel des IMPP. Es wurden keine Patientendaten verarbeitet; ein Ethikvotum war daher nicht erforderlich.

Einleitung

Künstliche Intelligenz (KI) hat sich innerhalb weniger Jahre von einer technologischen Nischenanwendung zu einem strukturbildenden Faktor zentraler gesellschaftlicher Funktionssysteme entwickelt. Besonders im Gesundheitswesen verändert der Einsatz großer Sprachmodelle nicht nur Entscheidungsprozesse, sondern auch Lern-, Ausbildungs- und Wissensstrukturen [1], [4]. Während sich ein Großteil der öffentlichen Debatte auf diagnostische oder therapeutische Anwendungen konzentriert, bleibt ein ebenso relevanter Bereich häufig unbeachtet: das medizinische Prüfungswesen als Schnittstelle zwischen Ausbildung und Versorgung.

Medizinische Staatsexamina erfüllen eine doppelte Funktion. Einerseits dienen sie der individuellen Leistungsbewertung, andererseits fungieren sie als institutionelles Steuerungsinstrument zur Sicherung von Qualität, Patientensicherheit und professionellen Standards im Sinne professioneller Governance [5]. Prüfungen definieren implizit, welches Wissen und welche Kompetenzen als berufsrelevant gelten. Wenn ein KI-System in der Lage ist, diese Prüfungen mit hoher Erfolgswahrscheinlichkeit zu bestehen, wirft dies weniger technische als vielmehr Fragen der Governance auf.

Ziel dieses Artikels ist es daher nicht, die Leistungsfähigkeit von KI isoliert zu bewerten, sondern deren Bedeutung für bestehende Prüfungs-, Ausbildungs- und Versorgungslogiken einzuordnen. Auf Basis einer empirischen Analyse von Prüfungsfragen des Instituts für medizinische und pharmazeutische Prüfungsfragen (IMPP) wird untersucht, inwieweit ein großes Sprachmodell standardisierte medizinische Wissensprüfungen bewältigen kann und welche Implikationen sich daraus für Entscheidungsträger ergeben.

Studie

Datengrundlage

Die empirische Analyse basiert auf einem Datensatz von insgesamt 1.252 Prüfungsfragen des Instituts für medizinische und pharmazeutische Prüfungsfragen aus dem Jahr 2024. Die Fragen wurden über die Lernplattform AMBOSS bezogen und umfassen sowohl vorklinische als auch klinische Inhalte der ärztlichen Staatsexamina in den Ausbildungsabschnitten M1 und M2 [7], [8]. Abgedeckt wurde eine breite Palette medizinischer Fachdisziplinen. Alle Fragen lagen im Single-Best-Answer-Format mit fünf Antwortoptionen (A–E) vor und entsprechen in Struktur und Anspruch den regulären schriftlichen Staatsexamina.

Public Management and Governance Review – 2026
DOI: https://doi.org/10.60733/PMGR.2026.02

Durchführung der Erhebung

Die Beantwortung der Prüfungsfragen erfolgte mithilfe des generativen KI-Modells Google Gemini 2.5 Pro (Web-Version, Stand Juli 2025) über die öffentlich zugängliche Web-Oberfläche. Die Datenerhebung wurde im Juli 2025 durchgeführt. Jede Frage wurde einzeln und in einer neuen, separaten Sitzung an das Modell übermittelt, um Kontexteinflüsse zwischen den Fragen zu vermeiden, analog zu etablierten Benchmark-Studien zu großen Sprachmodellen im medizinischen Kontext [5], [6].

Die Eingabe umfasste den vollständigen Fragentext, die Antwortoptionen sowie eine standardisierte Aufforderung, die korrekte Antwort (A–E) auszuwählen. Enthielten die Prüfungsfragen bildbasierte Inhalte, wurden diese zusätzlich in die Eingabe integriert. Die Antworten von Google Gemini wurden unmittelbar dokumentiert und mit dem offiziellen IMPP-Lösungsschlüssel abgeglichen.

Auswertung

Die Bewertung erfolgte strikt dichotom: Übereinstimmungen mit dem Lösungsschlüssel wurden als korrekt, Abweichungen als falsch klassifiziert. Die Auswertung beschränkt sich bewusst auf deskriptive Kennzahlen, insbesondere Trefferquoten insgesamt sowie differenziert nach Ausbildungsabschnitt (M1/M2) und Fächergruppen (vorklinisch/klinisch).

Ergebnisse

Die zentralen Ergebnisse sind in Tabelle 1 zusammengefasst. Über alle 1.252 untersuchten Prüfungsfragen hinweg erzielte Google Gemini eine Gesamt-Trefferquote von 96 Prozent. Damit wurde die übliche Bestehensgrenze medizinischer Staatsexamina von rund 60 Prozent deutlich überschritten [7].

Tabelle 1: Übersicht der Ergebnisse nach Ausbildungsabschnitt und Fächergruppe

Prüfungsteil	**Anzahl Fragen**	**Richtig**	**Trefferquote**
Vorklinische Fächer (M1)	787	754	96 %
Klinische Fächer (M1)	130	125	96 %
Vorklinische Fächer (M2)	9	7	78 %
Klinische Fächer (M2)	326	312	96 %
Gesamt	**1.252**	**1.198**	**96 %**

Note: Die geringere Trefferquote in den vorklinischen M2-Fächern ist aufgrund der sehr kleinen Fallzahl vorsichtig zu interpretieren

Vergleichbare Leistungsniveaus großer Sprachmodelle in medizinischen Prüfungsformaten wurden auch in anderen Studien berichtet, insbesondere bei standardisierten Multiple-Choice-Formaten [5], [6], [9].

Public Management and Governance Review – 2026
DOI: https://doi.org/10.60733/PMGR.2026.02

Implikationen für Entscheidungsträger

Die vorliegenden Ergebnisse begründen keinen unmittelbaren Handlungszwang für Prüfungsbehörden wie das Institut für medizinische und pharmazeutische Prüfungsfragen. Der Einsatz von KI-Systemen ist während der ärztlichen Staatsexamina untersagt, und es bestehen keine Hinweise darauf, dass diese Regelung aktuell systematisch unterlaufen wird. Aus formaler Sicht bleiben die Prüfungen damit regelkonform und valide.

Die Relevanz der Ergebnisse liegt vielmehr auf einer vorgelagerten Ebene. Große Sprachmodelle können in der Prüfungsvorbereitung ohne Einschränkung genutzt werden und sind faktisch bereits Teil des Lernökosystems vieler Studierender. Entsprechende Nutzungsmuster werden auch in der medizinischen Ausbildung international beschrieben [1], [2].

Darüber hinaus berühren die Ergebnisse auch die klinische Praxis. Die in medizinischen Prüfungen abgefragten Wissensbestände bilden zugleich die Grundlage vieler ärztlicher Routinetätigkeiten, etwa bei der Einordnung von Symptomen, der Auswahl leitlinienbasierter Therapieoptionen oder der Anwendung standardisierter Klassifikationssysteme. Entsprechend wird der Einsatz von KI-Systemen im klinischen Alltag zunehmend als unterstützende Informations- und Entscheidungsressource diskutiert [4].

Gleichzeitig verweisen zahlreiche Arbeiten auf die Grenzen dieser Systeme, insbesondere auf Halluzinationen, fehlendes Kontextverständnis und das Risiko überzeugend formulierter Fehlinformationen [3], [10]. Dies macht eine unreflektierte Delegation ärztlicher Verantwortung problematisch.

In diesem Kontext gewinnt das dritte Staatsexamen (M3) besondere Bedeutung. Als mündlich-praktische Prüfung adressiert es gezielt jene Kompetenzdimensionen, die sich einer Standardisierung entziehen, darunter klinisches Denken, ärztliche Kommunikation und Entscheidungsfindung unter Unsicherheit [7]. Die hohe Leistungsfähigkeit von KI in den schriftlichen Prüfungen M1 und M2 relativiert die Rolle des M3 nicht, sondern unterstreicht dessen Funktion als institutionelles Gegengewicht zu wissensbasierten Prüfungsformaten. Damit fungiert das M3 nicht nur als Abschlussprüfung, sondern als institutionelles Korrektiv innerhalb eines zunehmend KI-gestützten Wissensumfelds.

Empfehlungen

Die Empfehlungen sind bewusst auf strategischer Ebene formuliert und zielen nicht auf kurzfristige regulatorische Eingriffe ab, sondern auf eine mittelfristige Weiterentwicklung des Ausbildungs- und Versorgungssystems.

Die empirischen Befunde lassen sich in konkrete Handlungsempfehlungen übersetzen. Tabelle 2 fasst diese Empfehlungen entlang zentraler Governance-Akteure zusammen.

Public Management and Governance Review – 2026
DOI: https://doi.org/10.60733/PMGR.2026.02

Tabelle 2: Handlungsempfehlungen nach Akteursgruppe

Akteur	Zentrale Empfehlung	Ziel
Prüfungsbehörden (z.B. IMPP)	Systematische Analyse jener Fragetypen, die von KI besonders zuverlässig beantwortet werden	Sicherung der Prüfungsvalidität
Medizinische Fakultäten	Integration von KI-Kompetenz und kritischem Umgang mit LLMs in die Lehre	Professionalisierung der Ausbildung
Klinikleitungen	Klare Leitlinien zur unterstützenden, nicht delegierenden Nutzung von KI	Patientensicherheit
Gesundheitspolitik	Entwicklung eines Governance-Rahmens für KI-gestützte Entscheidungsunterstützung	Transparenz und Verantwortlichkeit
Ausbildungssystem gesamt	Stärkere Profilierung des M3 als Prüfung nicht delegierbarer Kompetenzen	Zukunftsfähigkeit des Berufsbilds

Diese Empfehlungen verdeutlichen, dass die Ergebnisse nicht auf eine punktuelle Anpassung einzelner Prüfungsformate zielen, sondern eine koordinierte Weiterentwicklung des gesamten Ausbildungs- und Versorgungssystems nahelegen.

Fazit

Die Ergebnisse dieses Artikels zeigen, dass Google Gemini medizinische Prüfungsfragen des IMPP mit einer sehr hohen Erfolgsquote beantworten kann. Daraus ergibt sich kein unmittelbarer Reformdruck für Prüfungsordnungen. Die eigentliche Bedeutung der Befunde liegt in ihrer langfristigen Relevanz für Ausbildungs- und Versorgungslogiken.

Wenn KI-Systeme in der Prüfungsvorbereitung und im klinischen Alltag zunehmend leistungsfähig werden, verändert sich das professionelle Wissensumfeld strukturell. Für Entscheidungsträger bedeutet dies nicht die Abwehr einer Technologie, sondern die aktive Gestaltung ihres Einsatzes. Das Zusammenspiel von M1, M2 und M3 bietet hierfür bereits ein differenziertes institutionelles Modell, das durch den gezielten Einsatz von KI nicht ersetzt, sondern strategisch weiterentwickelt werden kann.

Über den Autor

Thomas Gönner ist im österreichischen Gesundheitswesen tätig und verfügt über umfangreiche Erfahrung in der Konzeption, Implementierung und Weiterentwicklung patientenorientierter Informations- und Leitsysteme in komplexen Krankenhausumgebungen. Seine berufliche Tätigkeit ist in technischen Organisationsstrukturen im Krankenhausbetrieb verortet, die Planung, Koordination und systemische Steuerung verbinden. Er absolvierte den MBA in Health Care Management an der Wirtschaftsuniversität Wien. Seine inhaltlichen Schwerpunkte liegen in der Anwendung künstlicher Intelligenz im Gesundheitswesen, der medizinischen Ausbildung und dem Prüfungswesen sowie in Governance- und Entscheidungsfragen an der Schnittstelle von Technologie, Organisation und medizinischer Versorgung.

Quellenangaben

[1] Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023 Mar 19;11(6):887. doi: 10.3390/healthcare11060887. PMID: 36981544; PMCID: PMC10048148.

[2] Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023 Mar 6;9:e46885. doi: 10.2196/46885. PMID: 36863937; PMCID: PMC10028514.

[3] Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2024 Jul-Aug;17(5):926-931. doi: 10.1002/ase.2270. Epub 2023 Mar 28. Erratum in: Anat Sci Educ. 2024 Dec;17(9):1779. doi: 10.1002/ase.2496. PMID: 36916887

[4] Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023 Sep 5;330(9):866-869. doi: 10.1001/jama.2023.14217. PMID: 37548965.

[5] Jung, Leonard B. and Gudera, Jonas A. and Wiegand, Tim L. T. and Allmendinger, Simeon and Dimitriadis, Konstantinos and Koerte, Inga K. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. Dtsch Arztebl International, vol. 120, no. 21-22, pp. 373–374, 2023, doi: 10.3238/arztebl.m2023.0113.

[6] Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. JMIR Med Educ. 2024 Feb 8;10:e50965. doi: 10.2196/50965. PMID: 38329802; PMCID: PMC10884900.

[7] Bundesministerium für Gesundheit, “Approbationsordnung für Ärzte (ÄApprO 2002), §14 Abs. 6”.

[8] Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), “Bestehens- und Notengrenzen.”. Available: https://www.impp.de/pruefungen/allgemein/bestehens-und-notengrenzen.html. [Accessed: Mar. 14, 2026]

[9] Mihalache A, Grad J, Patil NS, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye (London, England). 2024 Sep;38(13):2530-2535. DOI: 10.1038/s41433-024-03067-4. PMID: 38615098; PMCID: PMC11383935.

[10] Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern Promethean dilemma. Croat Med J. 2023 Feb 28;64(1):1-3. doi: 10.3325/cmj.2023.64.1. PMID: 36864812; PMCID: PMC10028563.