Comparative analysis between the performance of ChatGPT and medical students outside their examination phase

Vergleichende Analyse der Leistungen von ChatGPT und Medizinstudenten außerhalb ihrer Prüfungsphase

Abstract

Introduction: State-run medical licensing examinations in Germany are highly demanding, requiring extensive preparation. While medical expertise remains central to licensure, patients increasingly seek advice from artificial intelligence (AI) models like ChatGPT. This shift raises the question of whether AI can accurately convey essential medical knowledge. Previous international studies (e.g., China, US, Poland, UK) have compared AI performance with that of certified professionals, often through indirect comparisons using historical exam averages. However, no studies have directly compared AI to medical students assessed outside their exam preparation phase, which could provide insight into knowledge retention. This study aims to directly compare the performance of large language models (LLMs) with that of medical students beyond their examination phase.

Methods: An anonymized survey was conducted at a German medical school among students in the clinical stage of their studies (typically 170 to 180 students). Participants answered 10 single-choice questions randomly selected from a pre-filtered pool derived from past German preclinical medical exam (M1) items. Questions were selected based on clinical relevance, moderate difficulty, and exclusion of chemical/mathematical content. The same questions were answered by ChatGPT-3.5, ChatGPT-4, and ChatGPT-4 mini. Performance was compared in terms of correct responses. Additionally, the corrected discrimination coefficient was calculated for each item, measuring how well each question differentiated between higher and lower performers.

Results: Of the 143 participants (median age 22), 129 were in the 5th semester, and the rest were in later semesters. About 40% identified as male and 55% as female. Students answered a median of 7 out of 10 questions correctly (range: 1–10). All AI models answered 9 out of 10 questions correctly. The only question missed by AI was answered correctly by 35% of students (50/143) and had the second-highest discrimination coefficient (0.28), indicating it effectively differentiated student performance.

Discussion: LLMs outperformed medical students who were beyond their exam preparation phase. However, limitations include the modest sample size and preselection of questions based on specific criteria (clinical relevance, moderate difficulty, exclusion of chemical/mathematical content), which introduces selection bias and limits the representativeness of our findings for the full examination. Importantly, the AI's one incorrect answer had the second highest, although still marginal discrimination coefficient, highlighting a possible gap in AI understanding for nuanced or complex content. These findings suggest that AI models are not without limitations and should be supplemented by human oversight, particularly in high-stakes or ambiguous clinical contexts.

Conclusion: AI models demonstrate strong performance in answering single choice questions, surpassing students outside active exam pre-

Daniel Leufkens[1]
Jörn Pons-Kühnemann[1]
Henning Schneider[1]
Anita C. Windhorst[1]

1 Institute of Medical Informatics, Department of Medicine, Justus Liebig University of Giessen, Germany

paration. However, their occasional errors, especially on discriminative questions, underline the need for caution. Further research is necessary to evaluate AI utility in real-world medical education and clinical decision-making, ensuring ethical and responsible integration.

Keywords: artificial intelligence, AI, large language models, LLMs, ChatGPT, performance on medical examination, comparative analysis, experiment

Zusammenfassung

Einleitung: Die staatlichen ärztlichen Zulassungsprüfungen in Deutschland sind sehr anspruchsvoll und erfordern eine umfangreiche Vorbereitung. Während medizinisches Fachwissen nach wie vor eine zentrale Rolle bei der Approbation spielt, suchen Patienten zunehmend Rat bei Modellen der künstlichen Intelligenz (KI) wie ChatGPT. Diese Entwicklung wirft die Frage auf, ob KI in der Lage ist, grundlegendes medizinisches Wissen korrekt zu vermitteln. Frühere internationale Studien (z.B. in China, den USA, Polen und dem Vereinigten Königreich) haben die Leistung von KI mit der von zertifizierten Fachleuten verglichen, häufig durch indirekte Vergleiche anhand historischer Prüfungsdurchschnitte. Es gibt jedoch keine Studien, in denen die Leistungen von KI direkt mit denen von Medizinstudenten verglichen wurden, die außerhalb ihrer Prüfungsvorbereitungsphase beurteilt wurden, was einen Einblick in das Abrufen von Wissen geben könnte. Ziel dieser Studie ist es, die Leistung von großen Sprachmodellen (LLMs) direkt mit der von Medizinstudenten außerhalb ihrer Prüfungsphase zu vergleichen.

Methoden: An einer deutschen medizinischen Hochschule wurde eine anonymisierte Umfrage unter Studierenden in der klinischen Phase ihres Studiums durchgeführt (in der Regel 170 bis 180 Studierende). Die Teilnehmer beantworteten 10 Single-Choice-Fragen, die nach dem Zufallsprinzip aus einem vorgefilterten Pool von Aufgaben aus dem 1. Abschnitt der in Deutschland zentral durchgeführten Ärztlichen Prüfung (M1) ausgewählt wurden. Die Fragen wurden nach klinischer Relevanz, mittlerem Schwierigkeitsgrad und Ausschluss von chemisch-mathematischen Inhalten ausgewählt. Die gleichen Fragen wurden von ChatGPT-3.5, ChatGPT-4 und ChatGPT-4 mini beantwortet. Die Leistung wurde anhand der Anzahl der richtigen Antworten verglichen. Zusätzlich wurde der korrigierte Trennschärfe-Koeffizient berechnet.

Ergebnisse: Von den 143 Teilnehmern (Durchschnittsalter 22) befanden sich 129 im 5. Semester, der Rest in späteren Semestern. Ca. 40% der Studierenden identifizierten sich als männlich, 55% als weiblich. Die Studierenden beantworteten im Median 7 von 10 Fragen richtig (Spanne: 1–10). Alle KI-Modelle beantworteten 9 von 10 Fragen richtig. Die einzige Frage, die von der KI falsch beantwortet wurde, wurde von 35% der Studierenden (50/143) richtig beantwortet und wies den zweithöchsten Trennschärfe-Koeffizienten (0,28) auf, was darauf hindeutet, dass sie die Leistungen der Studierenden effektiv differenziert.

Diskussion: Die LLMs schnitten besser ab als Medizinstudenten. Zu den Einschränkungen gehören jedoch die eher geringe Stichprobengröße und die Vorauswahl der Fragen nach spezifischen Kriterien (klinische Relevanz, mittlerer Schwierigkeitsgrad, Ausschluss chemisch-mathematischer Inhalte), was einen Selektionsbias einführt und die Repräsentativität unserer Befunde für die gesamte Prüfung einschränkt. Wichtig ist, dass die durch die LLMs falsch beantworte Frage eine, verglichen mit den anderen Fragen, hohe Trennschärfe aufwies, was auf eine mögliche Lücke im KI-Verständnis für nuancierte oder komplexe Inhalte hinweist. Diese Ergebnisse deuten darauf hin, dass KI-Modelle nicht uneingeschränkt einsetzbar sind und durch menschliche Aufsicht ergänzt werden sollten, insbesondere in anspruchsvollen oder mehrdeutigen klinischen Kontexten.

Schlussfolgerung: KI-Modelle übertreffen Studierende außerhalb der aktiven Prüfungsvorbereitung. Ihre gelegentlichen Fehler, insbesondere bei Fragen mit hoher Trennschärfe, machen jedoch deutlich, dass Vorsicht geboten ist. Weitere Forschung ist notwendig, um den Nutzen von KI in der realen medizinischen Ausbildung und im Studium zu bewerten.

Schlüsselwörter: künstliche Intelligenz, KI, Large Language Models, LLMs, ChatGPT, Leistung in medizinischen Prüfungen, vergleichende Analyse, Experiment

Introduction

State licensing examinations in medicine are highly demanding both in terms of subject matter and content, requiring months of preparation from candidates. Despite the rigorous medical expertise required for medical licensure maintaining its excellent reputation, patients are increasingly seeking medical advice from artificial intelligence (AI) with large language models such as ChatGPT. This paradigm shift raises fundamental questions about whether AI systems are capable of conveying necessary medical knowledge appropriately, both in terms of subject matter and content.

Consequently, several studies have compared results achieved by AI systems with those of certified medical examiners across various countries. Research has been conducted in China [1], the United States [2], [3], Poland [4], and the United Kingdom [5]. Multiple systematic reviews have synthesized these findings, with Brin et al. [3] reporting 80-90% accuracy for large language models on medical examinations, Liu et al. [6] analyzing ChatGPT performance across different versions worldwide, and Jin et al. [7] demonstrating an overall effect size of 70.1%, with 69.1% only in the field of medicine, in their meta-analysis. These studies typically conduct indirect comparisons with examination averages from corresponding examination years. However, direct comparison between AI performance and medical students assessed with temporal distance from their examination phase has not yet been investigated.

Recent studies have extended beyond basic licensing examinations to examine AI performance in specialized medical domains. Longwell et al. [8] evaluated large language models on medical oncology examination questions, finding 85% accuracy, though they noted that 81.8% of incorrect answers could lead to patient harm. Similarly, Tarabanis et al. [9] tested publicly available large language models on internal medicine board-style questions. These specialized assessments highlight the complexity of medical knowledge evaluation and the potential risks associated with AI-generated medical advice. While AI systems demonstrate consistent performance patterns without temporal degradation, human learners experience natural forgetting curves that affect knowledge retention over time [10]. E.g., in medical students, knowledge retention rates drop to around 53% to 70% in the area of physiology in the span of an average of 16 weeks [11]. Knowledge retention significantly depends on the learning method (e.g., active retrieval practice vs. passive review) [12], [13], the practical application of learned material in clinical or simulation contexts [14], the number of repetitions and testing frequency including overlearning [15], [16], and the spacing or timing between learning episodes and examinations (spacing effect) [13], [17]. Medical students who have completed their state examinations may demonstrate different performance patterns when assessed outside their active preparation phase, potentially providing insights into the practical implications of knowledge retention in medical practice. This distinction is particularly relevant when considering the real-world application of medical knowledge, where practitioners must recall information learned during their training years later in clinical practice.

Current research approaches have focused primarily on testing the maximum learning level of medical students during their examination preparation phase, rather than examining their later recall of previously learned material in comparison with AI systems. This methodological limitation restricts our understanding of how AI performance compares to the practical reality of medical knowledge application in clinical settings. Importantly, according to Miller's framework [18] of clinical competence, factual knowledge ("knows/knows how") represents only the foundational levels of competence, whereas clinical decision-making requires higher levels of performance in simulated or real contexts ("shows how/does") [18], [19]. Thus, the present study specifically targets the comparison of pre-clinical factual knowledge rather than clinical reasoning or decision-making skills.

Therefore, the aim of this study is to establish a direct comparison of pre-clinical knowledge levels between AI systems and medical students assessed outside their examination phase, using questions familiar to students from their previous state examination experience. This approach should provide a more nuanced understanding of AI capabilities in pre-clinical knowledge compared to traditional indirect comparisons with historical examination averages.

Methods

An anonymized survey was conducted among medical students at the medical faculty of the Justus-Liebig-University Gießen who were already in the clinical stage of their studies. In a typical semester around 170–180 students are enrolled. The survey required medical students to answer a random selection of single-choice examination questions familiar to them from their own state examination from the previous year in 2024 (first section of the German medical examination “M1”). Questions were selected by the authors from a pre-filtered pool based on clinical relevance (reference to diagnostics: Q01, 10, case studies: Q03, 07, 09 or diseases: Q02, 04, 05, 06, 08), moderate difficulty (according to Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP) [20], the average correct answer rate for the set of questions Q01–10 was 66.8% [Min. correct answer rate 27% for Q03 and max. 90% for Q06]), and exclusion of chemical and mathematical content to focus on clinically applicable medical knowledge, which has been taught at least up to the M1 level (see Table 1 for complete question set with translations). Students were surveyed during their clinical phase, creating temporal distance from their active examination preparation period.

The same questions were answered by ChatGPT-3.5, ChatGPT-4, and ChatGPT-4 mini to enable direct performance comparison.

Statistical analysis

Performance was evaluated by comparing correct response rates between AI models and medical students. A binomial test was used to determine if students answered randomly, success rate is then at 20% (one in five answers). Additionally, the corrected discrimination coefficient [21] was calculated for each question to measure how effectively each item differentiated between higher and lower-performing students, providing insight into question quality and the nature of knowledge being assessed. The corrected discrimination coefficient is a point-biserial correlation that quantifies the relationship between correctness on a single item and the overall test score. Based on established psychometric standards, we interpret the discrimination coefficients as follows: values $r \geq 0.4$ indicate very good discrimination, $0.3 \leq r < 0.4$ reasonably good discrimination, $0.2 \leq r < 0.3$ marginal discrimination, and $r < 0.2$ poor discrimination. Question 9 achieved a coefficient of $r=0.28$, placing it in the marginal discrimination category [22].

Statistical analysis was performed using R version 4.5.1 [23].

Results

Demographics of student cohort

A total of 143 students participated in the study with a median age of 22 years (mean = 23.04, SD=5.16, range: 19–69). The majority of participants were in the 5th semester (n=129, 90.2%) of the German medical studies system, having completed their first state-run examination at the end of the previous semester. The remaining participants were in later semesters (6th–23rd semester, mean = 5.30, SD=1.65).

Gender distribution was as follows: 57 participants (39.9%) identified as male, 79 participants (55.2%) as female, 2 participants as diverse, and 5 participants did not specify their gender.

Overall performance results

Students achieved a median score of 7 out of 10 questions correctly (mean = 6.77, SD=1.78, range: 1–10, interquartile range: 6–8), which is statistically different from randomly selecting one of five possible answers (probability for success: 0.7, 95% CI: 0.35–0.93, $p<0.001$). Students in the 5th semester showed a median of 7 of 10 correct answers (mean = 6.83, SD=1.74, range: 1–10, interquartile interval: 6–8), and students in higher semesters (n=13, 7 in the 6th, 2 in the 8th, and one in the 7th, 9th, 10th, and 23rd semester) answered a median of 6 out of 10 questions correctly (mean = 6.15, SD=2.15, range: 2–9, interquartile range: 5–7). Difference in total points were not statistically significant (Wilcoxon rank sum test, $p=0.243$).

The distribution showed a slight negative skew (–0.67), indicating that most students performed above the mean, with relatively few scoring very low (Figure 1). In contrast, all three LLM models (ChatGPT-3.5, ChatGPT-4, and ChatGPT-4 mini) achieved identical performance, answering 9 out of 10 questions correctly (90% accuracy).

Question difficulty and discrimination analysis

The analysis of individual questions revealed considerable variation in difficulty and discriminative power (Table 2). The questions ranged from very easy (Q01: 91.6% correct) to very difficult (Q03: 25.9% correct). Notably, the question that all LLMs answered incorrectly (Q09) had moderate difficulty (35% student success rate) and showed the second highest, though still marginal discrimination (0.28), indicating that it was one of the more effective questions at differentiating between higher and lower-performing students.

Table 1: Questions used and their translation

Q	Original	Translation
01	Die Sonografie ist ein bildgebendes Verfahren, das auf der Wechselwirkung von Schallwellen mit dem zu untersuchenden Gewebe basiert. Welchen Vorteil hat dabei die Verwendung höherer Schallfrequenzen im Vergleich zu niedrigeren Schallfrequenzen? (A) bessere Nutzung der transversalen Anteile der Schallwellen (B) Reduktion der Erwärmung des Gewebes bei der Untersuchung **(C) Verbesserung der Ortsauflösung bei der Bildgebung [88%]** (D) Verringerung der Ausbreitungsgeschwindigkeit (E) Verringerung der Totalreflexion der Schallwellen am Sendekopf	Sonography is an imaging technique based on the interaction of sound waves with the tissue being examined. What is the advantage of using higher sound frequencies compared to lower sound frequencies? (A) Better utilization of the transverse components of the sound waves (B) Reduction of tissue heating during the examination **(C) Improvement of spatial resolution in imaging [88%]** (D) Reduction of the propagation velocity (E) Reduction of total reflection of the sound waves at the transmitter head
02	Bei länger andauernder Nahrungskarenz wird der Stoffwechsel so umgeleitet, dass zur Aufrechterhaltung der Blutglucose-Konzentration verschiedene Organe bzw. Gewebe mit der Leber zusammenarbeiten. Wie trägt das weiße Fettgewebe am ehesten zu dieser Zusammenarbeit bei? Das weiße Fettgewebe ... (A) betreibt in geringem Ausmaß Gluconeogenese **(B) gibt Glycerin aus dem Abbau der Triacylglycerole (Triglyceride) an das Blut ab [81%]** (C) gibt in großen Mengen Alanin aus dem Proteinabbau an das Blut ab (D) sezerniert Adiponectin zur Steigerung der Gluconeogenese (E) synthetisiert Ketonkörper und gibt sie an das Blut ab	During prolonged fasting, the metabolism is redirected so that various organs and tissues work together with the liver to maintain blood glucose levels. How does white adipose tissue contribute most to this cooperation? White adipose tissue... (A) performs gluconeogenesis to a small extent **(B) releases glycerol from the breakdown of triacylglycerols (triglycerides) into the blood [81%]** (C) releases large amounts of alanine from protein breakdown into the blood (D) secretes adiponectin to increase gluconeogenesis (E) synthesizes ketone bodies and releases them into the blood
03	Bei einer 50-jährigen Patientin mit zunehmenden Muskel- und Knochenschmerzen sowie mehreren erlittenen Frakturen fallen niedrige Plasmakonzentrationen von Phosphat und Calcitriol auf. Als Ursache für diese Veränderungen wird ein endokrin aktiver Weichteiltumor identifiziert. Welches der folgenden Proteine produziert dieser Tumor am ehesten? (A) Calcitonin **(B) Fibroblast growth factor 23 (FGF23) [27%]** (C) Insulin-like growth factor (IGF-1) (D) Parathormon (E) Somatotropin	A 50-year-old female patient with increasing muscle and bone pain and multiple fractures has low plasma concentrations of phosphate and calcitriol. An endocrine-active soft tissue tumor is identified as the cause of these changes. Which of the following proteins is this tumor most likely to produce? (A) Calcitonin **(B) Fibroblast growth factor 23 (FGF23) [27%]** (C) Insulin-like growth factor (IGF-1) (D) Parathyroid hormone (E) Somatotropin
04	Der Vitamin-D-Status kann durch die Bestimmung der 25-Hydroxycholecalciferol-(Calcidiol-)Konzentration im Serum ermittelt werden. In welchem der folgend genannten Organe bzw. Gewebe wird 25-Hydroxycholecalciferol (Calcidiol) direkt gebildet? (A) Darm (B) Haut (C) Knochen **(D) Leber [68%]** (E) Niere	Vitamin D status can be determined by measuring the concentration of 25-hydroxycolecalciferol (calcidiol) in serum. In which of the following organs or tissues is 25-hydroxycolecalciferol (calcidiol) produced directly? (A) Intestines (B) Skin (C) Bones **(D) Liver [68%]** (E) Kidneys
05	Welche der folgend genannten Verbindungen stimuliert die Synthese von Aldosteron in den Zellen der Nebennierenrinde am stärksten? (A) adrenocorticotropes Hormon (ACTH) **(B) Angiotensin II [81%]** (C) Parathormon (D) Prostacyclin (E) Somatostatin	Which of the following compounds most strongly stimulates the synthesis of aldosterone in the cells of the adrenal cortex? (A) Adrenocorticotropic hormone (ACTH) **(B) Angiotensin II [81%]** (C) Parathyroid hormone (D) Prostacyclin (E) Somatostatin

(Continued)

Table 1: Questions used and their translation

Q	Original	Translation
06	Manche Hormone können aufgrund ihrer lipophilen Eigenschaften durch die Zellmembran diffundieren, anschließend an intrazelluläre Rezeptoren binden und diese aktivieren. Die Wirkung welches der folgend genannten Hormone wird auf diese Weise vermittelt? (A) Adrenalin (B) Calcitonin **(C) Estradiol [90%]** (D) Glucagon (E) Insulin	Due to their lipophilic properties, some hormones can diffuse through the cell membrane, bind to intracellular receptors, and activate them. Which of the following hormones acts in this way? (A) Adrenaline (B) Calcitonin **(C) Estradiol [90%]** (D) Glucagon (E) Insulin
07	Welcher der folgenden Parameter nimmt als Folge des physiologischen Alterungsvorgangs am ehesten zu? **(A) Durchmesser der Aorta [46%]** (B) Skelettmuskelmasse (C) T-Lymphozytenzahl (D) Tiefschlafphasen (E) Totalkapazität der Lunge	Which of the following parameters is most likely to increase as a result of the physiological aging process? **(A) Diameter of the aorta [46%]** (B) Skeletal muscle mass (C) T lymphocyte count (D) Deep sleep phases (E) Total lung capacity
08	Verschiedene Schmerzmittel hemmen Cyclooxygenasen. Auf welchen Prozess der nozizeptiven Reizaufnahme bzw. -weiterleitung wirken diese Schmerzmittel daher am ehesten? (A) deszendierende Bahnung **(B) periphere Sensibilisierung [74%]** (C) segmentale Hemmung (D) Wind-up-Phänomen (E) zentrale Sensibilisierung	Various painkillers inhibit cyclooxygenases. Which process of nociceptive stimulus reception or transmission are these painkillers most likely to affect? (A) Descending facilitation **(B) Peripheral sensitization [74%]** (C) Segmental inhibition (D) Wind-up phenomenon (E) Central sensitization
09	Bei einer 53-jährigen Patientin wird ein systemarterieller Blutdruck von 150/95 mmHg gemessen. Dieser erhöhte Blutdruck kann zu einer deutlichen Steigerung des Harnminutenvolumens im Vergleich zum Harnminutenvolumen bei normwertigem Blutdruck führen. Wodurch wird das Harnminutenvolumen in dieser Situation am ehesten gesteigert? (A) Abfall des kolloidosmotischen Drucks im Vas afferens der Niere (B) Anstieg der GFR in den kortikalen Glomeruli **(C) gesteigerte Durchblutung des Nierenmarks [36%]** (D) verstärkte Sekretion von Aldosteron ins Blutplasma (E) verstärkte Sekretion von Harnstoff im kortikalen Sammelrohr	A 53-year-old female patient has a systemic arterial blood pressure of 150/95 mmHg. This elevated blood pressure can lead to a significant increase in urine output compared to urine output at normal blood pressure. What is the most likely cause of the increase in urine output in this situation? (A) Decrease in colloid osmotic pressure in the afferent vessels of the kidney (B) Increase in GFR in the cortical glomeruli **(C) Increased blood flow to the renal medulla [36%]** (D) Increased secretion of aldosterone into the blood plasma (E) Increased secretion of urea in the cortical collecting duct
10	Im Arbeitsdiagramm des Muskels wird die Länge auf der x-Achse (Abszisse) und die Kraft auf der y-Achse (Ordinate) aufgetragen. Bei welcher Kontraktionsform entsteht hierbei immer eine vertikale Linie? (A) Anschlagszuckung (B) auxotone Kontraktion **(C) isometrische Kontraktion [77%]** (D) isotone Kontraktion (E) Unterstützungszuckung	In the working diagram of the muscle, length is plotted on the x-axis (abscissa) and force on the y-axis (ordinate). Which type of contraction always produces a vertical line? (A) Twitch contraction (B) Auxotonic contraction **(C) Isometric contraction [77%]** (D) Isotonic contraction (E) Support contraction

Notes: Correct answer is marked in bold; in square percentage answered correctly in M1 2024 according to the IMPP.

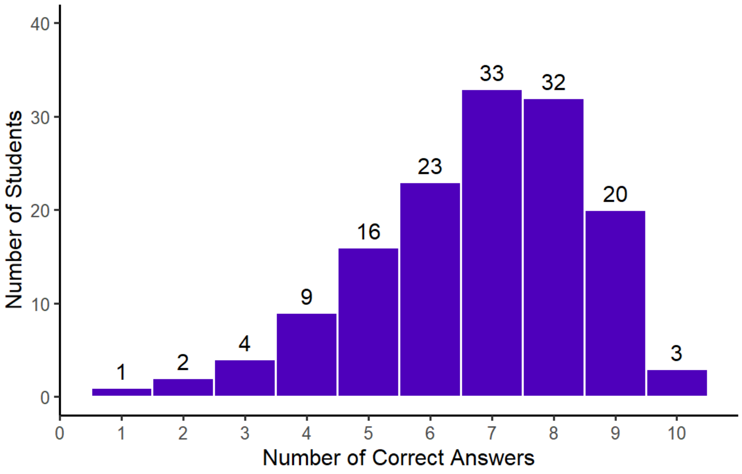

Figure 1: Distribution of total student scores

Table 2: Results of individual questions and AI performance

Question	Correct (n)	Difficulty (correct, %)	AI performance	Discrimi-nation
Q01	131	91.6%	Correct	0.15
Q02	121	84.6%	Correct	0.37
Q03	37	25.9%	Correct	0.26
Q04	95	66.4%	Correct	0.21
Q05	107	74.8%	Correct	0.24
Q06	135	94.4%	Correct	0.28
Q07	68	47.6%	Correct	0.07
Q08	111	77.6%	Correct	0.25
Q09	50	35.0%	Incorrect	0.28
Q10	113	79.0%	Correct	0.19

Analysis of AI failure case

The question that all LLMs failed to answer correctly comes from the case-related questions (Q03, 07, 09) and, according to IMPP [20], was answered correctly by 36% of M1 participants in 2024. It is important to note that item difficulty and discriminatory power represent distinct psychometric properties. The low proportion of correct responses (36%) reflects difficulty, while the discrimination coefficient of 0.28 reflects adequate discriminatory power, indicating that students with stronger overall performance were more likely to answer this question correctly. Thus, despite being challenging, this item effectively differentiated between performance levels, which is a desirable property for assessment items. This question required understanding of complex renal physiology and the relationship between systemic hypertension and renal hemodynamics, particularly the differential effects on cortical versus medullary blood flow:

Question Q09 (German): “Bei einer 53-jährigen Patientin wird ein systemarterieller Blutdruck von 150/95 mmHg gemessen. Dieser erhöhte Blutdruck kann zu einer deutlichen Steigerung des Harnminutenvolumens im Vergleich zum Harnminutenvolumen bei normwertigem Blutdruck führen. Wodurch wird das Harnminutenvolumen in dieser Situation am ehesten gesteigert?”

A – Abfall des kolloidosmotischen Drucks im Vas afferens der Niere
B – Anstieg der GFR in den kortikalen Glomeruli
C – gesteigerte Durchblutung des Nierenmarks [correct]
D – verstärkte Sekretion von Aldosteron ins Blutplasma
E – verstärkte Sekretion von Harnstoff im kortikalen Sammelrohr

Translation: “A 53-year-old patient has a systemic arterial blood pressure of 150/95 mmHg measured. This elevated blood pressure can lead to a significant increase in urine minute volume compared to urine minute volume at normal blood pressure. How is the urine minute volume most likely increased in this situation?”
A – Decrease in colloid osmotic pressure in the afferent arteriole of the kidney
B – Increase in GFR in the cortical glomeruli
C – Increased blood flow to the renal medulla [correct]
D – Increased secretion of aldosterone into the blood plasma
E – Increased secretion of urea in the cortical collecting duct

- AI answer: B – Anstieg der GFR in den kortikalen Glomeruli (Increase in GFR in cortical glomeruli), 64/143 of students also gave this answer, which is the majority of students for this question

- Correct answer: C – Gesteigerte Durchblutung des Nierenmarks (Increased blood flow to the renal medulla), student success rate: 50/143 (35.0%)

Discussion

Comparison with existing literature

The present study contributes to the growing body of literature examining AI performance in medical examinations by providing a direct comparison between AI systems and medical students assessed outside their examination phase. Previous research has established that AI language models can achieve substantial accuracy on medical licensing examinations. Our findings align with recent systematic reviews, where Brin et al. [3] reported 80-90% accuracy for large language models on medical examinations, and Jin et al. [7] demonstrated an overall effect size of 70.1% across multiple studies. The 90% median accuracy observed for LLMs in our study falls within the upper range of these reported values.

Country-specific studies have similarly demonstrated strong AI performance across diverse medical examination systems. The research conducted has shown that AI systems can achieve results comparable to or even superior to examination averages [1], [2], [4]. However, these studies predominantly rely on indirect comparisons with historical examination data rather than direct contemporaneous assessment. Our study addresses this methodological limitation by providing a direct comparison between AI and medical students, offering a more nuanced understanding of AI's capabilities in answering single-choice questions.

Performance analysis

The results indicate that LLMs achieved a median accuracy of 90%, outperforming medical students who were assessed after their exam preparation period and scored a median of 67.7%. Several factors must be considered when interpreting this difference. Students assessed outside of their active study period may exhibit reduced performance due to established forgetting curves associated with long-term memory decay [10]. In contrast, LLMs demonstrate consistent performance patterns without temporal degradation. The time interval of approximately 16 weeks since active preparation likely influenced student performance, as the retention of medical knowledge is known to decline in the absence of repeated reinforcement. The study we conducted took place approximately 26 weeks after the medical examination (October 2024 to April 2025). This interpretation aligns with previous findings; for example, a study reported a decrease in student accuracy from 70.4% to 53.5% in the topic of physiology over a similar duration [11].

The single question that challenged all AI models (Q09) had moderate difficulty for students (35% success rate) and although marginal, still the second highest discrimination (0.28), suggesting it tested nuanced understanding rather than factual recall. This finding indicates that the question differentiated relatively good between higher and lower-performing students while simultaneously exposing limitations in AI comprehension of complex physiological concepts.

While students showed variable performance across questions, AI models made a consistent error on the same complex physiology question, indicating potential systematic gaps in understanding rather than random errors typical of human performance.

Implications for medical education

The findings have several implications for medical education and clinical practice. AI models could serve as valuable resources for knowledge reinforcement and self-assessment, providing students with immediate feedback and comprehensive coverage of medical topics. The discriminative question that AI failed may represent areas where human clinical reasoning surpasses current AI capabilities, highlighting the complementary nature of human and artificial intelligence in medical contexts.

Furthermore, AI performance could help identify questions that effectively differentiate student competency levels, potentially informing assessment design and curriculum development. The consistent AI performance across different model versions suggests reliable knowledge base consistency, which could be valuable for standardized educational applications. However, this must be balanced against the risks identified in specialized medical domains. Longwell et al. [8] found that while AI achieved 85% accuracy on medical oncology examination questions, 81.8% of incorrect answers could lead to patient harm, indicating a significant risk profile for AI-generated medical advice.

The systematic nature of AI errors, as demonstrated by the consistent failure on complex physiology questions, suggests that AI limitations may be predictable and identifiable. This predictability could inform the development of hybrid educational approaches that leverage AI strengths while addressing its systematic weaknesses through human expertise.

Limitations and future directions

Several limitations should be acknowledged when interpreting these results. The study included 143 students from a single institution, which may limit the generalizability of findings to other medical schools or educational systems. The restricted set of 10 preselected questions constitutes the main limitation of this study. While items were chosen to represent clinically relevant, moderately difficult pre-clinical content, this purposive sampling strategy introduces selection bias. The deliberate exclusion of chemical/mathematical content and the focus on moderate difficulty levels mean that our findings cannot be generalized to the full spectrum of medical knowledge or difficulty ranges assessed in licensing examinations.

Consequently, the small number of questions prevents a systematic error analysis of where and why AI models fail. Our findings should therefore be interpreted as an exploratory pilot comparison rather than a definitive benchmark. Future studies should include larger and more diverse item pools covering multiple content domains and cognitive levels, which would enable more granular error analysis and increase the generalizability of results [19].

Moreover, the temporal distance between examination preparation and assessment may have differentially affected individual students, and the specific question selection may have influenced the observed performance gap. Additionally, the study focused on multiple-choice questions from a single examination system, which may limit generalizability to other assessment formats or medical education systems. While our study focused on factual knowledge, the evaluation of clinical decision-making requires a wide range of assessment approaches that extend beyond single-best-answer formats. Daniel et al. [19] have shown that multiple-choice questions can, under certain conditions, be applied to assess aspects of clinical reasoning such as leading diagnosis and treatment decisions; however, such question types were not included in our item set. Therefore, no inference regarding clinical competence or decision-making skills should be drawn from our results.

Future research should examine the stability of AI performance across different medical specialties and question formats, as well as investigate the factors that contribute to knowledge retrieval in both AI systems and human learners. Particular attention should be paid to identifying systematic patterns in AI errors and developing methods to address these limitations. The development of more sophisticated evaluation frameworks that account for the temporal dimension of medical knowledge retention would provide valuable insights for medical education and AI development.

Conclusion

This study provides the first direct comparison between AI performance and medical students assessed outside their examination phase, revealing a significant performance advantage for AI systems. While these findings suggest potential applications for AI in medical education, they must be interpreted within the context of the broader literature highlighting both the capabilities and limitations of AI in medical contexts. The superiority demonstrated by AI systems, combined with the identified risks in specialized medical domains, emphasizes the need for careful consideration of AI implementation in medical education and practice.

Notes

Author contributions

- Study conception: all authors
- Study realization: DL, ACW, JPK
- Drafting the manuscript: ACW, DL
- Revising the manuscript: all authors
- Data analysis: ACW, DL

Authors’ ORCIDs

- Daniel Leufkens: 0000-0002-9729-2905
- Jörn Pons-Kühnemann: 0000-0002-8211-4399
- Henning Schneider: 0000-0002-9958-4434
- Anita C. Windhorst: 0000-0002-7357-2080

Competing interests

The authors declare that they have no competing interests.

References

1. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024 Feb 14;24(1):143. DOI: 10.1186/s12909-024-05125-7
2. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312. DOI: 10.2196/45312
3. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How GPT models perform on the United States medical licensing examination: a systematic review. Discov Appl Sci. 2024;6(10):500. DOI: 10.1007/s42452-024-06194-5
4. Suwała S, Szulc P, Guzowski C, Kamińska B, Dorobiała J, Wojciechowska K, Berska M, Kubicka O, Kosturkiewicz O, Kosztulska B, Rajewska A, Junik R. ChatGPT-3.5 passes Poland’s medical final examination-Is it possible for ChatGPT to become a doctor in Poland? SAGE Open Med. 2024 Jun 17;12:20503121241257777. DOI: 10.1177/20503121241257777
5. Vij O, Calver H, Myall N, Dey M, Kouranloo K. Evaluating the competency of ChatGPT in MRCP Part 1 and a systematic literature review of its capabilities in postgraduate medical assessments. PLoS One. 2024 Jul 31;19(7):e0307372. DOI: 10.1371/journal.pone.0307372
6. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024 Jul 25;26:e60807. DOI: 10.2196/60807
7. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. 2024 Sep 16;24(1):1013. DOI: 10.1186/s12909-024-05944-8

8. Longwell JB, Hirsch I, Binder F, Gonzalez Conchas GA, Mau D, Jang R, Krishnan RG, Grant RC. Performance of Large Language Models on Medical Oncology Examination Questions. JAMA Netw Open. 2024 Jun 3;7(6):e2417641. DOI: 10.1001/jamanetworkopen.2024.17641

9. Tarabanis C, Zahid S, Mamalis M, Zhang K, Kalampokis E, Jankelson L. Performance of Publicly Available Large Language Models on Internal Medicine Board-style Questions. PLOS Digit Health. 2024 Sep 17;3(9):e0000604. DOI: 10.1371/journal.pdig.0000604

10. Ebbinghaus H. Memory: a contribution to experimental psychology. Ann Neurosci. 2013 Oct;20(4):155-6. DOI: 10.5214/ans.0972.7531.200408

11. Csaba G, Szabó I, Környei JL, Kerényi M, Füzesi Z, Csathó Á. Variability in knowledge retention of medical students: repeated and recently learned basic science topics. BMC Med Educ. 2025 Apr 11;25(1):523. DOI: 10.1186/s12909-025-07096-9

12. Anders ME, Vuk J, Rhee SW. Interactive retrieval practice in renal physiology improves performance on customized National Board of Medical Examiners examination of medical students. Adv Physiol Educ. 2022 Mar 1;46(1):35-40. DOI: 10.1152/advan.00118.2021

13. Deng F, Gluckstein JA, Larsen DP. Student-directed retrieval practice is a predictor of medical licensing examination performance. Perspect Med Educ. 2015 Dec;4(6):308-13. DOI: 10.1007/s40037-015-0220-x

14. Larsen DP, Dornan T. Quizzes and conversations: exploring the role of retrieval in medical education. Med Educ. 2013 Dec;47(12):1236-41. DOI: 10.1111/medu.12274

15. Fraundorf SH, Caddick ZA, Nokes-Malach TJ, Rottman BM. Cognitive perspectives on maintaining physicians' medical expertise: III. Strengths and weaknesses of self-assessment. Cogn Res Princ Implic. 2023 Aug 30;8(1):58. DOI: 10.1186/s41235-023-00511-z

16. Kornell N, Hays MJ, Bjork RA. Unsuccessful retrieval attempts enhance subsequent learning. J Exp Psychol Learn Mem Cogn. 2009;35(4):989-98. DOI: 10.1037/a0015729

17. Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychol Bull. 2006 May;132(3):354-80. DOI: 10.1037/0033-2909.132.3.354

18. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990 Sep;65(9 Suppl):S63-7. DOI: 10.1097/00001888-199009000-00045

19. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, Estrada CA, Ballard T, Artino AR Jr, Sergio Da Silva A, Cleary T, Stojan J, Gruppen LD. Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med. 2019 Jun;94(6):902-12. DOI: 10.1097/ACM.0000000000002618

20. AMBOSS SE. Physikum 2024. [cited 2025 Jul 14]. Available from: https://next.amboss.com/de/questions

21. Bortz J, Döring N. Forschungsmethoden und Evaluation. 4. ed. Berlin, Heidelberg: Springer; 2006. Quantitative Methoden der Datenerhebung; p. 137-293. DOI: 10.1007/978-3-540-33306-7_4

22. Zubairi NA, AlAhmadi TS, Ibrahim MH, Hegazi MA, Gadi FU. Effective use of Item Analysis to improve the Reliability and Validity of Undergraduate Medical Examinations: Evaluating the same exam over many years: a different approach. Pak J Med Sci. 2025 Mar;41(3):810-5. DOI: 10.12669/pjms.41.3.10693

23. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2025.

Corresponding author:

Dr. Anita C. Windhorst
Institut für Medizinische Informatik, Abt. Medizinische Statistik, Rudolf-Buchheim-Str. 6, 35392 Gießen, Germany, Phone: +49 641 99 41366
anita.c.windhorst@informatik.med.uni-giessen.de

Please cite as

Leufkens D, Pons-Kühnemann J, Schneider H, Windhorst AC. Comparative analysis between the performance of ChatGPT and medical students outside their examination phase. GMS Med Inform Biom Epidemiol. 2026;22:Doc05.
DOI: 10.3205/mibe000303, URN: urn:nbn:de:0183-mibe0003033

This article is freely available from

https://doi.org/10.3205/mibe000303

Published: *2026-03-25*