CORRESPONDENCE

Research Letter

ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted

Full medical registration in Germany requires passing three medical state examinations. The first state exam (M1) regarding pre-clinical subjects consists of a written exam and a viva voce exam. The second state exam (M2) is a written test and comprises the medical specialties. The third state exam (M3)is an oral/viva voce practical exam. The written test questions are published by the German Institute for Medical and Pharmaceutical Examination Questions (IMPP).Both written tests consist of 320 single choice questions with five possible answers each.

Recently, artificial intelligence (AI) in the form of ChatGPT passed examinations of the US Medical License exam (USMLE) (1). ChatGPT is a large language model (LLM), which is based on the transformer network architecture "general pre-trained transformer" (GPT) with more than 170 billion parameters (2, 3). It recognizes speech patterns and responds in a context appropriate way to user queries/questions.

We investigated whether Chat GPT could pass the first and second state examinations in Germany and thus is able to answer complex medical questions in the German language.

Methods

Our analyses are based on the questions of the written tests for the first (23–24 August 2022) and second (11–13 October 2022 state examinations, which were retrieved from the learning platform AMBOSS (www.amboss.com, AMBOSS GmbH, 6 March 2023). Questions that were not considered by the IMPP in the evaluation were excluded here too (M1: n=11; M2: n=9). ChatGPT does not allow for images to be input, and questions that required looking at images to give the answer were also excluded (image questions: M1: n=46; M2: n=59). Altogether 263 questions remained for M1 and 252 questions for M2. Each question was assigned to a test subject. In M2, distinction was made between case based questions (n=175) and non-case based questions (n=77). We used Chat GPT based on GPT 3.5 (https://chat.openai.com/; version: 13 February 2023; OpenAI). Examination questions can be accessed only on paid-for platforms. Furthermore, no exam question using identical wording existed in the five years before the training of the ChatGPT version used in our study had been concluded at the end of 2021. The probability is therefore high that the algorithm was unfamiliar with the exam questions studied here.

The questions were entered into ChatGPT and the response given by the algorithm was compared with the sample solutions in the examinations. Chi square tests were used to determine differences in the performance of ChatGPT between specialties and to compare the results in case based and non-case based questions. Spearman correlation analyses were used to determine correlations between the performance of ChatGPT and the percentage of students who ticked the correct answers to the questions in AMBOSS.

TABLE 1

Performance of ChatGPT in the first state examination (M1)

Subjects	Result (n, %, [95% CI])		
Anatomy	56	46.4	[33.0; 60.0]
Biochemistry	56	57.1	[43.8; 70.5]
Biology	18	77.8	[56.5; 99.1]
Chemistry	9	33.3	[0; 71.8]
Physics	11	45.5	[10.4; 80.5]
Physiology	54	63.0	[49.7; 76.3]
Psychology	30	73.3	[56.5; 90.1]
Sociology	29	75.9	[59.3; 92.4]
Total	263	60.1	[54.1; 65.8]

CI, confidence interval

Results

ChatGPT gave the correct answers in M1 for 60.1% (158/263) and in M2 for 66.7% (168/252) and thus passed both examinations with a grade 4 (pass grade, "sufficient"). For M1, differences were seen between test subjects (p=0.024; *Table 1*). In M1, the best results were achieved for biology (77.8%; 14/18), sociology (75.9%; 22/29), and psychology (73.3%; 22/30). ChatGPT achieved poorer results for chemistry (33.3%, 3/9), physics (45.5%; 5/11), and anatomy (46.4%; 26/56).

For M2 too, differences were seen between test subjects (p=0.045; *Table 2*). The best results were achieved in pharmacology (94.7%; 18/19), ophthalmology ((85.7%; 6/7), and dermatology (87.7%; 6/7). The worst results were seen for otorhinolaryngology (33.3%; 1/3), neurology (46.7%; 21/45), and epidemiology (46.7%; 7/15). No differences were seen (p=0.629) between the results for case based questions (65.7%; 115/175; 95% confidence interval [58.6; 72.8]) and non-case based questions (68.8 %; 53/77; [58.3; 79.4]).

The performance of ChatGPT correlated weakly with the percentage of students who ticked the correct question online (M1: $\rho = 0.207$; $p < 0.001$; [0.085; 0.323]; M2: $\rho = 0.288$; $p < 0.001$; [0.167; 0.400]).

Discussion

The LLM ChatGPT passed the written tests for M1 and M2 narrowly when image questions were excluded. It achieved a similar performance as in the US exams (1). ChatGPT therefore delivered a poorer overall performance than the average exam participants (students: M1=73.0%; M2=74.2%) (4).One explanation may be the fact that the medical questions were originally entered in German, whereas ChatGPT was 93% trained with English-language texts and without a medical focus (5).

TABLE 2

Performance of ChatGPT in the second state examination (M2)

Subjects	Result (n, %, [95% CI])		
AIEP	8	75.0	[36.3; 100]
Ophthalmology	7	85.7	[50.8; 100]
Surgery/orthopedics	14	64.3	[35.6; 93.0]
Dermatology	7	85.7	[50.8; 100]
Epidemiology	15	46.7	[18.1; 75.3]
Gynecology	20	80.0	[60.8; 99.2]
Otorhinolaryngology	3	33.3	[0; 100]
Human genetics	14	64.3	[35.6; 93.0]
Infectious diseases	13	84.6	[61.9; 100]
Internal medicine	34	64.7	[47.8; 81.6]
Neurology	45	46.7	[31.5; 61.8]
Pediatrics	23	65.2	[44.2; 86.3]
Pharmacology	19	94.7	[83.7; 100]
Psychiatry	9	66.7	[28.2; 100]
Radiology	8	75.0	[36.3; 100]
Forensic medicine	12	66.7	[35.4; 98.0]
Urology	1	100.0	/
Total	252	66.7	[60.6; 72.2]

AIEP, anesthesia, intensive care, emergency medicine, pain management/palliative care; CI, confidence interval

The extent to which ChatGPT gave the correct answer to a question correlated weakly with the results achieved by medical students when answering the same question. The different results achieved by ChatGPT in the different subject areas might be explained with the complexity of the questions and the available training dates. Questions that required an understanding of positional relation, multimodal diagnostics, or transfer knowledge often prompted poorer answers. Question that required calculations or formula conversions also led to ChatGPT failure in more instances. By comparison, questions regarding terminology/definitions in psychology and sociology were often answered correctly. The excellent result for pharmacology could be explained with the structured, freely available drug information. Future studies should investigate the performance of AI applications for image questions in relation to different question types.

Conclusions

These initial results show the performance of ChatGPT in giving answers to complex medical questions, based on the medical state examinations. The ability of LLMs to structure medical data and interpret information on the background of the available literature has potential in terms of the potential use of ChatGPT in medicine.

Leonard B. Jung*, Jonas A. Gudera*, Tim L. T. Wiegand*, Simeon Allmendinger, Konstantinos Dimitriadis, Inga K. Koerte
* The authors share joint first authorship.

cBRAIN, Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, Ludwig-Maximilians-Universität München, München (Jung, Wiegand, Koerte)
leonard.jung@gmx.com

Psychiatry Neuroimaging Laboratory, Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA (Jung, Wiegand, Koerte)

LMU AIM, Ludwig-Maximilians-Universität, München (Gudera, Wiegand)

Dr. von Hauner Children's Hospital, Ludwig-Maximilians-Universität, München (Gudera)

Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Harvard Medical School, Boston, MA, USA (Gudera)

Karlsruhe Institute of Technology, Karlsruhe (Allmendinger)

Neurological Clinic and Policlinic, Großhadern Hospital, Ludwig-Maximilians-Universität, München (Dimitriadis)

Institute for Stroke and Dementia Research (ISD), Ludwig-Maximilians-Universität, München

Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität, München (Koerte)

Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA (Koerte)

Conflict of interest statement
The authors declare that no conflict of interest exists.

Manuscript received on 14 March 2023, revised version accepted on 25 April 2023.

Translated from the original German by Birte Twisselmann, PhD.

References

1. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198.
2. Ouyang L, Wu J, Jiang X, et al.: Training language models to follow instructions with human feedback. NeurIPS 2022; 35: 27730–44.
3. OpenAI: Introducing ChatGPT. https://openai.com/blog/chatgpt (last accessed on 4 March 2023).
4. IMPP: Prüfungen Medizin – Lösungen und Ergbnisse. www.impp.de/pruefungen/medizin/l%C3%B6sungen-und-ergebnisse.html (last accessed on 4 March 2023).
5. Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. NeurIPS 2020; 33: 1877–901.

Cite this as:
Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK: ChatGPT passes German state examination in medicine with picture questions omitted. Dtsch Arztebl Int 2023; 120: 373–4. DOI: 10.3238/arztebl.m2023.0113