PeerJ Computer Science

Testing the knowledge of artificial intelligence chatbots in pharmacology: examples of two groups of drugs

Marcin Mateusz Granat, Aleksandra Paź and Dagmara Mirowska-Guzel

Department of Clinical and Experimental Pharmacology, Medical University of Warsaw, Warsaw, Poland

ABSTRACT

Objectives: The study aimed to evaluate eight artificial intelligence chatbots (ChatGPT-3.5, Microsoft Copilot, Gemini, You.com, Perplexity, Character.ai, Claude 3.5, and ChatRTX) in answering questions related to two pharmacological topics taught during the basic pharmacology curriculum for medical students: antifungal drugs and hypolipidemic drugs.

Methods: Chatbots' performance was assessed by answering 60 single-choice questions on antifungal and hypolipidemic drugs topics. The questions were designed to have four answers (a, b, c, and d), and the artificial intelligence (AI) role was to choose the proper one. The assessment was performed twice with a 1-year hiatus to determine if artificial intelligence chatbots' effectiveness changed over time. All the answers were checked for being right or wrong according to up-to-date pharmacology knowledge. To improve the clarity of results, to each score, a mark was assigned based on the grading system applied in our unit. Statistica software version 13.3 and Microsoft Excel 2010 were used for statistical analysis.

Results: In 2023, the best results on the subject of antifungal drugs were obtained by Gemini (formerly Bard) and on the topic of hypolipidemic drugs by You.com (formerly YouChat). In 2024 Microsoft Copilot answered correctly the highest number of questions in both topics. The total results of all artificial intelligence chatbots in 2023 and 2024 were compared using t-test for dependent samples. Statistical analysis revealed that artificial intelligence chatbots improved over time in both pharmacological topics, but this change was not statistically significant ($p = 0.784$ for antifungal drugs subject and $p = 0.056$ for hypolipidemic drugs).

Conclusions: The accuracy of AI chatbots' responses regarding antifungal and hypolipidemic drugs improved over one year, though not significantly. None of the tested AI systems provided correct answers to all questions within these pharmacological fields.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Computational Linguistics, Natural Language and Speech, World Wide Web and Web Science

Keywords Artificial intelligence, AI, AI chatbots, Pharmacology, Antifungal drugs, Hypolipidemic drugs

Submitted 14 February 2025
Accepted 22 May 2025
Published 15 July 2025

Corresponding author
Dagmara Mirowska-Guzel, dagmara.mirowska-guzel@wum.edu.pl

Academic editor
Luigi Di Biasi

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj-cs.2954**

Copyright
2025 Granat et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

In 1943 began the construction of the first electronic and programmable computer called the Electronic Numerical Integrator and Computer (ENIAC). At that moment question arose whether machines are intelligent (*Kissinger, Schmidt & Huttenlocher, 2022*). In

1950 mathematician and code breaker Alan Turing proposed an answer to this problem claiming that a machine should be considered "intelligent" if its behaviour cannot be distinguished from that of a human (*Turing, 1950*). This approach was labelled as the Turing test, and while its literal interpretation would allow passing only robots that are identical to humans, the pragmatic approach would rather call "intelligent" all machines that act in a human-like manner. According to that, chatbots can be classified as artificial intelligence (AI) systems based on their ability to mimic human conversation (*Kissinger, Schmidt & Huttenlocher, 2022*; *Tam et al., 2023*).

Chatbot is a computer program and a human-computer interaction (HCI, technology allowing user to interact with a computer *via* a specific interface) model that simulates communication by text or sound with human users, especially over the Internet (*Adamopoulou & Moussiades, 2020*). The AI chatbot's field is under constant development and refinement. Nowadays, this type of AI system already nearly matches or even exceeds human performance in tasks like, *e.g.*, competition-level mathematics, reading comprehension, and image classification (*Jones, 2024*). Chatbots are present in an enormous number of human activities. They are promising tools in medical education and the improvement of patient communication (*Sallam, 2023*). Chatbots can also generate queries that lead to high-precision searches, so their value for researchers who conduct systematic reviews is recognised (*Wang et al., 2023*). They have applications in customer service as an alternative to frequently asked questions (FAQ) by providing detailed replies (*Nirala, Singh & Purani, 2022*; *Kovacevic et al., 2024*).

Finally, chatbots are perceived as practical tools with future potential in treatment in many medical fields (*e.g.*, psychiatry, critical care nephrology, and urology) (*Cheng et al., 2023*; *Suppadungsuk et al., 2023*; *Talyshinskii et al., 2024*). A notable example is DiabeTalk, the AI chatbot with the ability to employ natural language understanding and decision-making algorithms to predict diabetes type and respond to user enquiries (*Rossi et al., 2024*). Another illustrious example of AI utilisation in medicine was presented in research by *De Roberto et al. (2024)*. In this study, ChatGPT-4o was analysing electrocardiogram images with the goal of assisting in the diagnosis of cardiovascular conditions.

As of today, chatbots have the ability to write referenced essays, and in the future, they may replace or be integrated into all search engines (*Stokel-Walker, 2022*, *2023*). Companies that develop AI do not always publish analyses on testing their systems. While the United States Food and Drug Administration (FDA) has approved hundreds of medical devices with AI implemented in healthcare facilities, between the years 2020 and 2022, only 65 randomized controlled trials of AI interventions were published (*Lenharo, 2024*; *Martindale et al., 2024*).

Pharmacology is one of an extremely fast-developing branch of medicine (*Brown et al., 2022*). Thousands of clinical trials are conducted each year worldwide, and huge number of clinical recommendations incorporating new data on drug application appear (*European Medicines Agency, 2024*; *Seoane-Vazquez, Rodriguez-Monguio & Powers, 2024*). Additionally, old drugs gain new indications, and drug repurposing is one of the most popular tools in the modern discovery of new therapies (*Pushpakom et al., 2019*). There is a

need for pharmacological researchers and industry to work together to develop and implement AI technologies, as well as an urgency to benchmark AI performance as an educational tool (*Aziz et al., 2024*; *Shahin et al., 2025*). Because there is limited data about chatbots' performance in the pharmacology field, this study's objective is to provide that evaluation.

MATERIALS AND METHODS

The aim of the study was to determine the ability of eight chatbots to correctly solve single-choice questions related to pharmacological knowledge on two groups of drugs: antifungal and hypolipidemic. Subsequently, based on the number of correctly answered questions, the comparison of chatbots was conducted. What is more, possessing knowledge that AI systems are constantly changing (*Tebenkov & Prokhorov, 2021*), the aforementioned questions were presented to chatbots twice with a 1-year break. The aim of that was to investigate if any alterations in their performance could be noticed.

In this study, Statistica software version 13.3 and Microsoft Excel 2010 were used for statistical analysis. A 95% confidence interval of the difference was applied for all statistical tests, so $p < 0.05$ was considered statistically significant. Evaluation if data are distributed normally was conducted using Lilliefors test and Shapiro–Wilk test. To assess AI performance before and after 1-year period (from August 2023 to August 2024), a t-test for dependent samples was applied. To highlight performance between AI chatbots analysis of variance (ANOVA) was conducted. For p values with borderline statistical significance threshold Cohen's d value was calculated to address effect size between groups.

Chatbots selection

AI chatbots were chosen on July 2, 2023, using search engines like Google, DuckDuckGo, and Bing, including phrases: "artificial intelligence chatbot", "artificial intelligence chat", and "AI chatbot". The second search with identical proceeding was performed on July 15, 2024, with the aim of finding new AI chatbots. In this study, we focused on AI systems that represent the large language models (LLMs) as they have already been widely deployed in medicine (*Thirunavukarasu et al., 2023*).

The large language model is an AI model that analyses text present in a vast number of books, articles, and internet-based content. This analysis is possible by using deep neural networks (computing systems inspired by biological neural networks with the ability to perform transformations upon input data), which allows AI to learn specific relationships between words. LLMs chatbots are able to generate responses to given tasks, but their accuracy and coherence are not always correct (*Thirunavukarasu et al., 2023*).

In this study, we focused on free-of-charge AI chatbots accessible *via* the internet and one AI chatbot implemented on GeForce™ RTX 4090 (Nvidia, USA) graphics processing unit (GPU) that we acquired on August 2, 2024.

Questions and their application

To determine AI chatbots' performance in the pharmacology field, 60 single-choice questions were prepared with four answers marked with letters a, b, c, and d. A total of 30

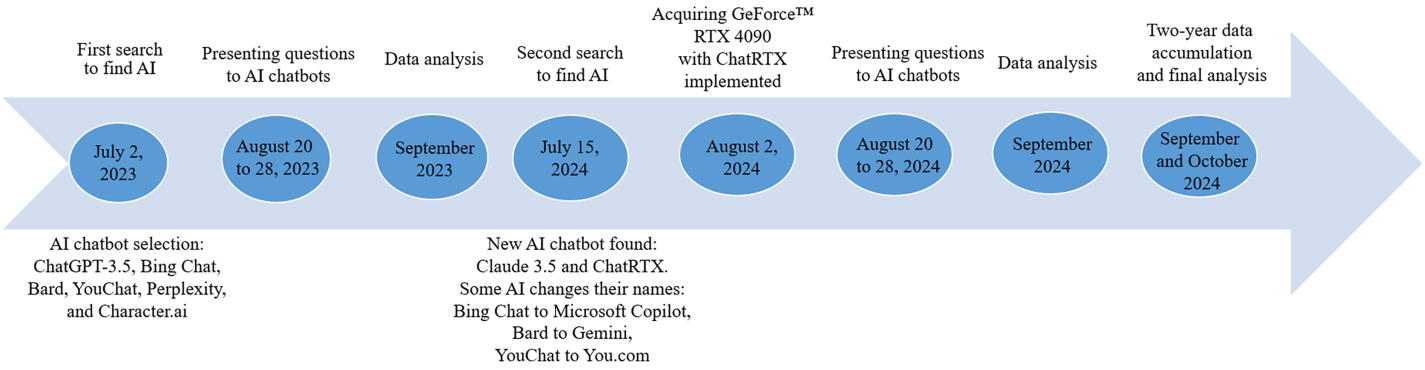

Figure 1 **Scheme presenting steps in testing AI chatbots. AI, artificial intelligence.** Full-size DOI: 10.7717/peerj-cs.2954/fig-1

questions were related to antifungal drugs and the other 30 to hypolipidemic drugs. All questions were designed to meet the university level of third-year medical students and were previously used in our recent study, which evaluated the effectiveness of the digital educational game in the pharmacology teaching process (*Granat, Paź & Mirowska-Guzel, 2024*). Three pharmacology experts prepared all questions to ensure the same level of difficulty.

On July 2, 2023, six AI chatbots were found: ChatGPT-3.5 (OpenAI, San Francisco, California, USA), Bing Chat (Microsoft, Redmond, WA, USA), Bard (Google AI, Mountain View, CA, USA), YouChat (You.com, Palo Alto, CA, USA), Perplexity (Perplexity AI, San Francisco, CA, USA), and Character.ai (Character Technologies, Menlo Park, CA, USA). From August 20 to 28, 2023 each chatbot was asked: "I would like to present to you a single-choice test questions about antifungal drugs. Your job will be to choose one correct answer. Is it all right with you?" After the declaration of agreement, 30 questions pertaining to antifungal drugs were presented to each AI chatbot. Subsequently, chatbots were asked a second question: "I would like to present to you a single-choice test questions about hypolipidemic drugs. Your job will be to choose one correct answer. Is it all right with you?", and after consent, 30 single-choice questions about hypolipidemic drugs were asked. All questions were presented one after the other, and in the same order, so chatbots were always working on one task at a time. All the answers were checked for being right or wrong according to current pharmacology knowledge.

On July 15, 2024, two new AI chatbots were found: Claude 3.5 (Anthropic, San Francisco, CA, USA), accessible *via* the internet and ChatRTX (Nvidia, Santa Clara, CA, USA), implemented on NVIDIA GeForce™ RTX 30 or 40 Series GPU or NVIDIA RTX™ Ampere or Ada Generation GPU with at least 8 GB of video random access memory (VRAM) (*NVIDIA, 2025*). At that time, we also noticed that some of the chatbots that we previously used had changed their names. Bing Chat was renamed Microsoft Copilot, Bard became Gemini, and YouChat turned into You.com. ChatGPT-3.5, Perplexity, and Character.ai remained under their original labels. From August 20 to 28, 2024, the same procedure of presenting questions as a year before was employed for all tested AI chatbots (see Fig. 1).

RESULTS

First comparison of AI chatbots

First testing took place from August 20 to 28, 2023 and included six AI chatbots: ChatGPT-3.5, Bing Chat, Bard, YouChat, Perplexity, and Character.ai. All AI systems declared that they can answer pharmacological questions. On the subject of antifungal drugs, the best result was achieved by Bard that answered correctly 26 questions out of all 30 (86.7%). This outcome was followed by Bing Chat (24 correct answers, 80%), YouChat (23 correct answers, 76.7%), Character.ai (22 correct answers, 73.3%), Perplexity (20 correct answers, 66.7%), and ChatGPT-3.5 (19 correct answers, 63.3%).

The application of 30 questions related to hypolipidemic drugs concluded that YouChat provided 27 correct answers (90%), which was the highest number among all tested chatbots. Next in line was Bard (26 correct answers, 86.7%), Bing Chat and ChatGPT-3.5 (both with 23 correct answers, 76.7%), Perplexity (22 correct answers, 73.3%), and Character.ai (21 correct answers, 70%).

The analysis of the answers on the subject of antifungal drugs revealed that all six AI chatbots answered correctly questions nos. 1, 2, 3, 8, 11, 16, 17, 18, 19, 21, 22, and 25. Questions nos. 5 and 9 were answered correctly by two AIs; one AI was right in question no. 28, and none provided proper solution in questions nos. 6 and 24. On the subject of hypolipidemic drugs, all six AI systems properly answered questions nos. 3, 6, 8, 10, 11, 12, 13, 16, 20, 21, 22, 23, and 24. Two AI chatbots answered correctly question no. 18 and one AI was right in questions nos. 19 and 29.

Second comparison of AI chatbots

After a hiatus of 1 year, from August 20 and 28, 2024, eight AI chatbots were employed for analysis. While ChatGPT-3.5, Microsoft Copilot, Gemini, You.com, Perplexity, Character.ai, and Claude 3.5 declared that would answer the questions, ChatRTX stated that it cannot provide any answers related to drugs or medical treatments. On this account, ChatRTX was excluded from this research.

The best result regarding antifungal drugs was achieved by Microsoft Copilot with 27 correct answers to 30 questions (90%). It was followed by Claude 3.5 (26 correct answers, 86.7%), You.com (25 correct answers, 83.3%), ChatGPT-3.5 and Perplexity (both with 23 correct answers, 76.7%), Character.ai (20 correct answers, 66.7%), and Gemini (19 correct answers, 63.3%). To improve the clarity of results, to each score, a mark was assigned based on the grading system applied in our university in which: a failing grade (2) was received when a number of correctly answered questions $(x) \in \langle 0,\ 15\rangle$, a satisfactory grade (3) when $x\ \in \langle 16, 18\rangle$, a satisfactory plus grade (3.5) when $x\ \in \langle 19,\ 21\rangle$, a good grade (4) when $x\ \in \langle 22,\ 24\rangle$, a good plus grade (4.5) when $x\ \in \langle 25,\ 27\rangle$, and an excellent grade (5) when $x\ \in \langle 28,\ 30\rangle$ (see Fig. 2).

Similarly, in the case of hypolipidemic drugs, Microsoft Copilot was also the best by answering 29 questions correctly (96.7%). Next were: Claude 3.5, Perplexity, and You.com,

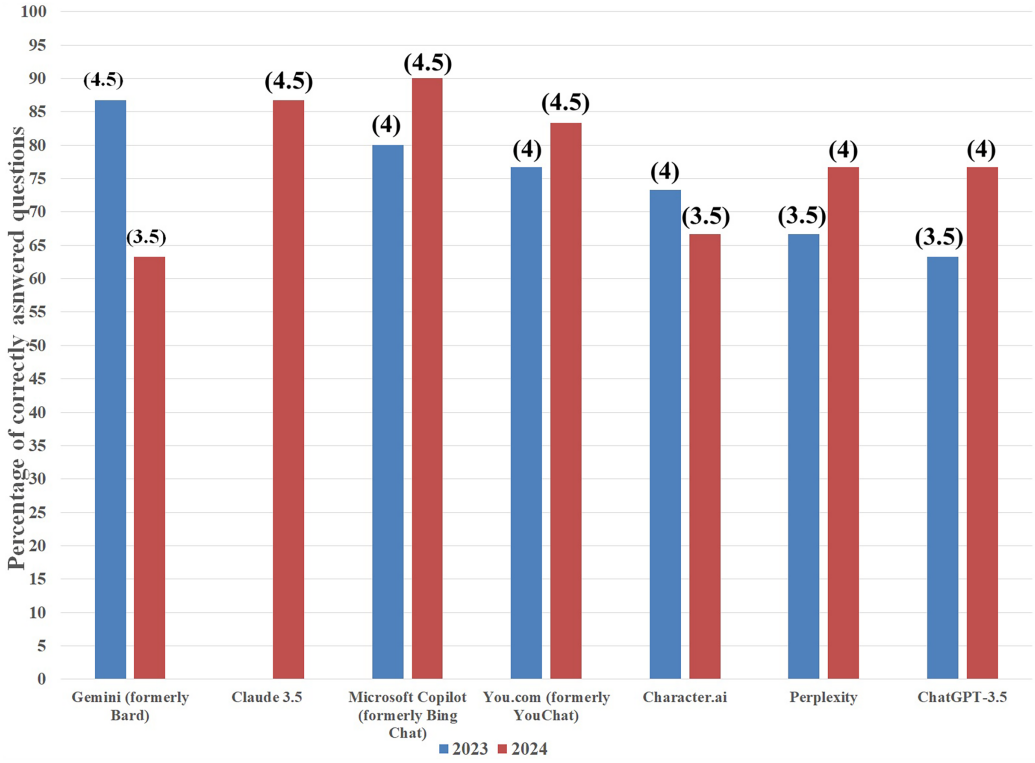

Figure 2 Artificial intelligence chatbots' percentage of correct answers on antifungal drugs topic in 2023 (blue columns) and 2024 (red columns). Grades attributed to each score are presented in parentheses.
Full-size DOI: 10.7717/peerj-cs.2954/fig-2

and all three successfully answered 28 questions (93.3%). ChatGPT-3.5 was next in line with 27 correct answers (90%), and after it were Character.ai (25 correct answers, 83.3%) and Gemini (24 correct answers, 80%). Identically, as with the results obtained in 2023, to each score the grade was assigned (see Fig. 3 and Table 1).

The answers on the subject of antifungal drugs were analysed, which revealed that questions nos. 2, 3, 4, 7, 8, 11, 13, 14, 16, 18, 19, 22, 25, 26, and 27 were properly answered by all seven AI chatbots. Questions nos. 15 and 24 were correctly resolved by two AIs, and questions nos. 5, 6, and 28 were properly solved by one AI. Analogical analysis on answers related to hypolipidemic drugs provided information that all seven AI systems were right in questions nos. 2, 3, 6, 8, 9, 10, 11, 12, 13, 16, 17, 21, 22, 23, 24, 25, 26, 27, 28, and 30. Only one AI answered correctly question no. 29.

Comparison of total results on AI chatbots performance

As all data obtained in this study represent normal distribution, the calculation of median or arithmetic mean was decided to be sufficient in the preliminary analysis of the results (*Gandhi et al., 2023*). In this study, we choose the median as a worth calculating value. According to that, the performance of all AI chatbots employed in the 2023 analysis characterised by median equalled 22.5 on the subject of antifungal drugs and 23.0 in relation to hypolipidemic drugs. The same calculation applied to all AI chatbots utilised in

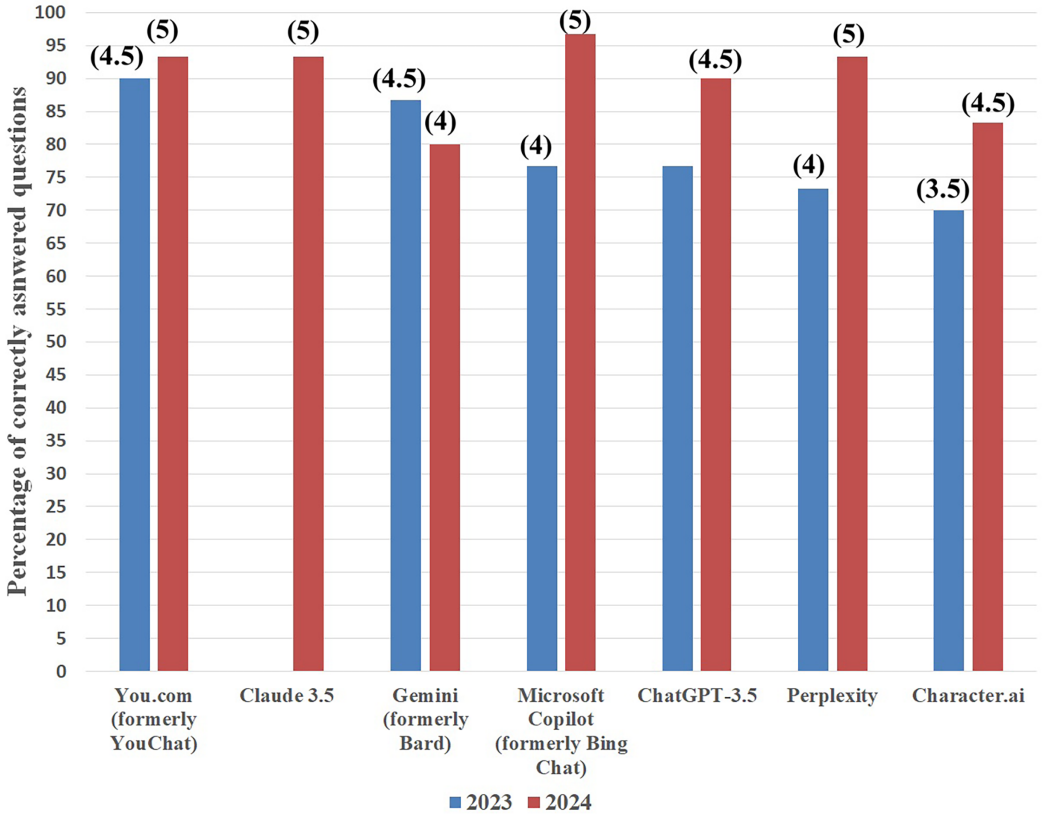

Figure 3 Artificial intelligence chatbots' percentage of correct answers on hypolipidemic drugs topic in 2023 (blue columns) and 2024 (red columns). Grades attributed to each score are presented in parentheses. Full-size DOI: 10.7717/peerj-cs.2954/fig-3

Table 1 Artificial intelligence chatbots' accuracy.

Artificial intelligence chatbot	Accuracy on antifungal topic		Accuracy on hypolipidemic topic	
	2023	2024	2023	2024
Gemini (formerly Bard)	86.7%	63.3%	86.7%	80%
Claude 3.5	–	86.7%	–	93.3%
Microsoft Copilot (formerly Bing Chat)	80%	90%	76.7%	96.7%
You.com (formerly You Chat)	76.7%	83.3%	90%	93.3%
Character.ai	73.3%	66.7%	70%	83.3%
Perplexity	66.7%	76.7%	73.3%	93.3%
ChatGPT-3.5	63.3%	76.7%	76.7%	90%

the 2024 analysis resulted in a median equalled 23.0 on the topic of antifungal drugs and 28.0 on the subject of hypolipidemic drugs.

Statistical analysis was conducted to determine if AI improvement over a year is statistically significant. Data sets distribution was analysed with the use of Lilliefors test and Shapiro–Wilk test and resulted in conclusion that all the data used in this study have normal distribution.

T-test for dependent samples conducted for antifungal drugs topic resulted in the conclusion that AI chatbots improved their performance, but this change was not statistically significant ($p = 0.784$). Likewise, the t-test for dependent samples employed to analyse data on the subject of hypolipidemic drugs stated that the AI betterment in answering questions was present but was not statistically significant ($p = 0.056$). Analysis of variance (ANOVA) on antifungal and hypolipidemic drugs subjects also resulted in no statistically significant changes with p values, respectively, $p = 0.183$ and $p = 0.860$. Cohen's d value for hypolipidemic drugs equalled 1.61.

All data on antifungal and hypolipidemic drugs together were analysed in relation to 2023 and 2024. The data represented a normal distribution, with median equalled 23.0 in 2023 and 25.5 in 2024. A t-test for dependent samples concluded that improvement of AI chatbots was not statistically significant ($p = 0.124$). Interestingly, all questions applied in this work were previously used in our different study on human participants (*Granat, Paź & Mirowska-Guzel, 2024*). Because of that AI systems results can be preliminarily compared with mean scores of third-year medical students ($n = 66$), which equalled 16.3 on antifungal drugs topic and 16.8 on hypolipidemic drugs topic.

DISCUSSION

All seven AI chatbots correctly answered most of the questions concerning antifungal and hypolipidemic drugs. Albeit the number of correct answers varied depending on the AI, none of them was inerrable. This is the reason why we conclude that the process of gaining knowledge in pharmacology field is still more beneficial using scientific sources rather than depending completely on AI systems.

Albeit all prepared questions were designed to represent the same level of difficulty, we observed that some of them were correctly answered by all AI chatbots, and others by only two, one, or no AI. No specific patterns (*e.g.*, questions categories) explaining AI performance were noticed. We cannot be certain, but AI difficulties may be a result of poor training data quality which led to inaccurate predictions of some answers (*Stanton, 2025*).

It is apparent that only one AI chatbot, Gemini (formerly Bard), has worsened its scores over time in both pharmacological topics. Although drops in AI performance and behaviour drifts (*e.g.*, changes in following user instructions) have been reported in the past, the causes behind them remain unclear (*Chen, Zaharia & Zou, 2024*). Companies developing AI systems are constantly updating their models, but specifics behind this process remain confidential. Because of that, the end user can only assess results of the updates, but is unable to identify a specific cause or causes of these results at *e.g.*, coding level. Although certainty eludes us, some propositions may suggest explanation. One of them is related to updates. They goal is to improve AI, but due to profound changes, they may result in the opposite effect (*OpenAI, 2023*). This situation may have happened with Gemini that received a major update on 8 February 2024 (*Gemini, 2024*), which was between first and second comparison of AI chatbots in this study. It is possible that these changes were caused by model collapse, which is a degenerative process in which data generated by AI end up polluting the training set of the next generation of generative

model. In effect AI misapprehends reality because is trained on a polluted collection of data (*Shumailov et al., 2024*).

It is important to emphasise that comparison of results on AI chatbots' performance on hypolipidemic drugs resulted in a borderline statistical significance threshold ($p = 0.056$) and Cohen's d value equalled 1.61, indicating that the effect size is large (*Cohen, 1988*). While it is true that AIs improved on the hypolipidemic drugs topic (median equalled 23.0 in 2023 and 28.0 in 2024), one must note that these results are based on aggregated data of seven AIs, whereby practical use of these results for a potential user is limited. It is so, because for an individual the most valuable information is which specific AI represents the best performance.

It is worth noting that the answers we obtained from AI chatbots differed from those of humans. The vast volume of text was generated within seconds in each answer, and presented language was grammatically accurate. It was in opposition to human responses, which are often concise, full of poor vocabulary, typographic errors, and abbreviations (*Hill, Ford & Farreras, 2015*). In our study, we also noticed that AI systems never used emojis, which are popular in human messages. It is evident that in our study a specific human-AI interaction was formed. Although our goal was not to determine the aforementioned interaction, future studies may take that under consideration and provide information on the dynamics of human cooperation with AI, which can be detected using *e.g.*, deep learning (*Freire-Obregón et al., 2020*).

The idea of testing AI systems, whether they can pass medical tests, has already been present in published studies. *Kung et al. (2023)* designed a study in which AI tried to pass the United States Medical Licensing Exam that was consisted of three standardized tests of expert-level knowledge. While AI employed in this research performed at or near the passing threshold of 60%, it is worth mentioning that only one type of AI, ChatGPT, was tested. Another approach, but also related to ChatGPT exclusively, was published by *Peng et al. (2024)*. In this work researchers did not use medical test but applied 131 valid questions from a medical book concerning colorectal cancer. The results stated that ChatGPT did not meet the standards of an expert level.

The assessment of medical knowledge performance related to more than one AI chatbot was conducted by *Pan et al. (2023)*. In this work, a queries related to the five most common cancers were presented to four AI systems, which resulted in the conclusion that the quality of text responses was good and no misinformation was present. A similar study was published by *Mohammad-Rahimi et al. (2024)*. Its methodology was based on an analysis of responses provided by six AI chatbots to questions related to oral pathology, oral medicine, and oral radiology with the use of a 5-point Likert scale. While the highest mean score for performance across all disciplines was 4.066 ± 0.825, the authors additionally evaluated the authenticity of citations generated by AI systems. Interestingly, 82 out of 349 (23.50%) citations were fake, which draws attention to a phenomenon called artificial hallucination—a situation when AI generates sensory experiences that appear real but are fictitious (*Tangsrivimol et al., 2025*). Although it was outside the scope of our study, future research may benefit from determining not only AI chatbots' effectiveness but also the level of their artificial hallucination.

As AI chatbots are becoming more and more popular in medical education, new fields of medicine become present in AI evaluation processes *e.g.*, family medicine in *Hanna et al. (2024)* study. In this work, researchers inputted 193 multiple-choice questions from the family medicine in-training exam written by the American Board of Family Medicine. Three AI chatbots were tested, and the best one scored 167/193 (86.5%), which was higher than residents' mean of 68.4%. Multiple-choice questions were also used in the testing of three AI systems on the subject of postgraduate-level orthopaedics (*Vaishya et al., 2024*). The AI with the highest score had 100% efficiency, but the study did not present information about the results obtained in humans on the same questions.

There is little data on the comparison of AI chatbots performance at answering medical questions in different time intervals. During our research, we found only one study of this type by *Mihalache, Popovic & Muni (2023)*, and it assessed ophthalmic knowledge of ChatGPT from January 9 to 16, 2023, and on February 17, 2023. The research concluded that ChatGPT answered half of the questions correctly, but its limitations were employing only one AI chatbot as well as short hiatus between AI assessments. Studies determining AI systems' performance regarding pharmacology knowledge alone are also very limited. We found one research by *Elango et al. (2023)* who tested ChatGPT abilities in pharmacology examination of phase II Bachelor of Medicine, Bachelor of Surgery (MBBS, a medical degree granted by universities in countries that adhere to the United Kingdom's higher education tradition) with the average total score results of 76%.

To the best of our knowledge, this study is the first assessment of numerous AI chatbots' performance in the pharmacology field with their comparison at two time intervals, applying a 1-year hiatus between them. Certainly, our work has its weaknesses. It is limited to only two pharmacological topics, so it would be beneficial to conduct future analysis with a broader scope of knowledge to enhance the generalizability of the findings. It is possible that AI chatbots' abilities will advance or regress in the course of time. With new versions of tested AI systems, examining their future performance would be worth considering for future research. What is more, as a field of artificial intelligence is steadily developing, it would be valuable to search for potential new AI systems and provide results on their effectiveness. It is worth noting that AI systems are exposed to data biases that may lead to skewed or unfair outcomes (*Hasanzadeh et al., 2025*). We would like to point out that presenting the same set of questions in the same order might introduce pattern recognition benefits for AI chatbots, so it may be beneficial for future studies to consider randomizing question order. What is more, AI chatbots can be trained on question-answer datasets that resemble the test format, which may affect AI responses (*Prakash et al., 2025*).

During all our work, we communicated with AI chatbots exclusively in English, but as large language models (LLMs) can learn other languages, future assessments of AI performance in different languages may be interesting to investigate. Moreover, additional insights on AI chatbots' performance in comparison to humans' abilities in executing the same tasks, would also be of great relevance and may be provided by prospective studies.

CONCLUSION

The novelty of this study includes a comparison of multiple AI chatbots' performance exclusively in pharmacology over a 1-year period. Moreover, some of the AIs (You.com, Perplexity, Character.ai, and ChatRTX) were never tested in the field of medicine knowledge, let alone pharmacology. The methodology of this work allowed us to determine AI chatbots' effectiveness in two pharmacologic topics, which is a useful insight as AI may be used as an educational tool.

The evaluation of eight AI chatbots resulted in the conclusion that in 2023, six of them answered correctly more than half if the questions (the range of correct answers for the antifungal drugs topic was from 19 to 26, and for the hypolipidemic drugs subject from 21 to 27). Better, but not with statistically significant change, were the results obtained in 2024 when seven AI chatbots were analysed. The range of correct answers for the antifungal drugs subject spanned from 19 to 27, and for the hypolipidemic drugs topic from 24 to 29. It is noticeable that only one AI chatbot, Gemini (formerly Bard), performed worse after a 1-year hiatus in both topics. What is more, ChatRTX was the only AI excluded from this study by being unable to answer enquiries.

As many AI chatbots are available free-of-charge, accessing them is instant and effortless. With rising number of users, it seems appropriate to seek data on AI chatbots' accuracy in delivered answers. While companies developing AI chatbots are not providing transparency tests (*Stokel-Walker, 2023*), determining AI knowledge by *e.g.*, employing methods presented in this study, may be the only tool suitable for verifying AI chatbots' usefulness as a source of information.

One may presume that the performance of AI chatbots in the field of pharmacology will improve over time but such an assumption cannot be fully justified. In our study, one of the evaluated AI chatbots surprisingly worsened within 1 year, whereas others improved, but none had reached a maximal score. It is essential information for students and teachers but also health professionals. They should be vigilant as the results of our study show that depending on AI chatbots may be deceptive.

ABBREVIATIONS

AI	artificial intelligence
ENIAC	Electronic Numerical Integrator and Computer
FAQ	frequently asked questions
FDA	Food and Drug Administration
GPU	graphics processing unit
HCI	Human-computer Interaction
LLMs	large language models
MBBS	Bachelor of Medicine, Bachelor of Surgery
VRAM	video random access memory

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors report there are no competing interests to declare.

Author Contributions

- Marcin Mateusz Granat conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Aleksandra Paź analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Dagmara Mirowska-Guzel analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available in the Supplemental Files and Zenodo: Granat, M. M. (2025). Testing the knowledge of artificial intelligence chatbots in pharmacology Examples of two groups of drugs. Zenodo. https://doi.org/10.5281/zenodo.15285002.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2954#supplemental-information.

REFERENCES

Adamopoulou E, Moussiades L. 2020. Chatbots: history, technology, and applications. *Machine Learning with Applications* **2(53)**:1–18 DOI 10.1016/j.mlwa.2020.100006.

Aziz MHA, Rowe C, Southwood R, Nogid A, Berman S, Gustafson K. 2024. A scoping review of artificial intelligence within pharmacy education. *American Journal of Pharmaceutical Education* **88(1)**:100615 DOI 10.1016/j.ajpe.2023.100615.

Brown DG, Wobst HJ, Kapoor A, Kenna LA, Southall N. 2022. Clinical development times for innovative drugs. *Nature Reviews Drug Discovery* **21**:793–794 DOI 10.1038/d41573-021-00190-9.

Chen L, Zaharia M, Zou J. 2024. How is ChatGPT's behavior changing over time? *Harvard Data Science Review* **6(2)**:1–26 DOI 10.1162/99608f92.5317da47.

Cheng SW, Chang CW, Chang WJ, Wang HW, Liang CS, Kishimoto T, Chang JP, Kuo JS, Su KP. 2023. The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and Clinical Neurosciences* **77(11)**:592–596 DOI 10.1111/pcn.13588.

Cohen J. 1988. *Statistical power analysis for the behavioral sciences*. New York: Routledge Academic.

De Roberto AM, De Marco F, Di Biasi L, Rossi D, Tortora G. 2024. Can ChatGPT-4o enhance ECG interpretation accuracy compared to cardiologists? In: *2024 IEEE International Conference*

on Bioinformatics and Biomedicine. Lisbon, Portugal, 6852–6858 DOI 10.1109/BIBM62325.2024.10822822.

Elango A, Kannan N, Anandan I, Surapaneni KM. 2023. Testing the knowledge and interpretation skills of ChatGPT in pharmacology examination of phase II MBBS. *Indian Journal of Pharmacology* **55(4)**:266–267 DOI 10.4103/ijp.ijp_188_23.

European Medicines Agency. 2024. Clinical trials in human medicines. *Available at https://www.ema.europa.eu/en/human-regulatory-overview/research-development/clinical-trials-human-medicines* (accessed 24 April 2025).

Freire-Obregón D, Castrillón-Santana M, Barra P, Bisogni C, Nappi M. 2020. An attention recurrent model for human cooperation detection. *Computer Vision and Image Understanding* **197–198(2)**:102991 DOI 10.1016/j.cviu.2020.102991.

Gandhi AK, Mishra V, Vashistha B, Rastogi M, Vashistha R. 2023. Distribution. In: Eltorai AEM, Bakal JA, Kim DW, Wazer DE, eds. *Handbook for Designing and Conducting Clinical and Translational Research, Translational Radiation Oncology*. First Edition. Cambridge: Academic Press, 131–133.

Gemini. 2024. Release updates. *Available at https://gemini.google.com/updates* (accessed 24 April 2025).

Granat MM, Paź A, Mirowska-Guzel D. 2024. The evaluation of digital educational game use in pharmacology teaching process. *Pharmacology Research & Perspectives* **12(5)**:1–8 DOI 10.1002/prp2.1237.

Hanna RE, Smith LR, Mhaskar R, Hanna K. 2024. Performance of language models on the family medicine in-training exam. *Family Medicine* **56(9)**:555–560 DOI 10.22454/FamMed.2024.233738.

Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. 2025. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *npj Digital Medicine* **8(154)** DOI 10.1038/s41746-025-01503-7.

Hill J, Ford WR, Farreras IG. 2015. Real conversations with artificial intelligence: a comparison between human-human online conversations and human-chatbot conversations. *Computers in Human Behavior* **49(3)**:245–250 DOI 10.1016/j.chb.2015.02.026.

Jones N. 2024. AI now beats humans at basic tasks—new benchmarks are needed, says major report. *Nature* **628**:8009 DOI 10.1038/d41586-024-01087-4.

Kissinger HA, Schmidt E, Huttenlocher D. 2022. *The Age of AI*. London: John Murray (Publishers).

Kovacevic S, Popovic T, Jovovic I, Cakic S, Babic D. 2024. Hotel chatbot receptionist for smart hospitality. In: *2024 28th International Conference on Information Technology (IT)*. Zabljak, Montenegro, 1–4.

Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2(2)**:1–12 DOI 10.1371/journal.pdig.0000198.

Lenharo M. 2024. The testing of AI in medicine is a mess. Here's how it should be done. *Nature* **632(8026)**:722–724 DOI 10.1038/d41586-024-02675-0.

Martindale APL, Llewellyn CD, de Visser RO, Ng B, Ngai V, Kale AU, di Ruffano LF, Golub RM, Collins GS, Moher D, McCradden MD, Oakden-Rayner L, Rivera SC, Calvert M, Kelly CJ, Lee CS, Yau C, Chan AW, Keane PA, Beam AL, Denniston AK, Liu X. 2024. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI

reporting guidelines. *Nature Communications* **15(1)**:1619 Erratum in: 2024. Nature Communications, 15: 6376 DOI 10.1038/s41467-024-45355-3.

Mihalache A, Popovic MM, Muni RH. 2023. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmology* **141(6)**:589–597 DOI 10.1001/jamaophthalmol.2023.1144.

Mohammad-Rahimi H, Khoury ZH, Alamdari MI, Rokhshad R, Motie P, Parsa A, Tavares T, Sciubba JJ, Price JB, Sultan AS. 2024. Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology* **137(5)**:508–514 DOI 10.1016/j.oooo.2024.01.015.

Nirala KK, Singh NK, Purani VS. 2022. A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications* **81(16)**:22215–22246 DOI 10.1007/s11042-021-11458-y.

NVIDIA. 2025. NVIDIA ChatRTX. Your personalized AI chatbot. *Available at https://www.nvidia.com/en-us/ai-on-rtx/chatrtx/* (accessed 24 April 2025).

OpenAI. 2023. Function calling and other API updates. *Available at https://openai.com/index/function-calling-and-other-api-updates/* (accessed 24 April 2025).

Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. 2023. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncology* **9(10)**:1437–1440 DOI 10.1001/jamaoncol.2023.2947.

Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, Zhang Y, Tang J, Gu Y, Sun Y. 2024. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. *Scientific Reports—Nature* **14(1)**:2840 DOI 10.1038/s41598-024-52853-3.

Prakash S, Cheng A, Yik J, Tschand A, Ghosal R, Uchendu I, Quaye J, Ma J, Grampurohit S, Giannuzzi S, Balyan A, Amin F, Pipersenia A, Choudhary Y. 2025. QuArch: a question-answering dataset for AI agents in computer architecture. *IEEE Computer Architecture Letters* **24(1)**:105–108 DOI 10.1109/LCA.2025.3541961.

Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M. 2019. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery* **18(1)**:41–58 DOI 10.1038/nrd.2018.168.

Rossi D, Citarella AA, De Marco D, Di Biasi L, Tortora G. 2024. Comparative analysis of diabetes diagnosis: WE-LSTM networks and WizardLM-powered DiabeTalk chatbot. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine*. Lisbon, Portugal, 6859–6866 DOI 10.1109/BIBM62325.2024.10821742.

Sallam M. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* **11(6)**:887 DOI 10.3390/healthcare11060887.

Seoane-Vazquez E, Rodriguez-Monguio R, Powers JH 3rd. 2024. Analysis of US food and drug administration new drug and biologic approvals, regulatory pathways, and review times, 1980–2022. *Scientific Reports–Nature* **14(1)**:3325 DOI 10.1038/s41598-024-53554-7.

Shahin MH, Desai P, Terranova N, Guan Y, Helikar T, Lobentanzer S, Liu Q, Lu J, Madhavan S, Mo G, Musuamba FT, Podichetty JT, Shen J, Xie L, Wiens M, Musante CJ. 2025. AI-driven applications in clinical pharmacology and translational science: insights from the ASCPT, 2024 AI preconference. *Clinical and Translational Science* **18(4)**:e70203 DOI 10.1111/cts.70203.

Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y. 2024. AI models collapse when trained on recursively generated data. *Nature* **631(8022)**:755–759 DOI 10.1038/s41586-024-07566-y.

Stanton D. 2025. Ten common mistakes to avoid when using AI in projects. In: *Project Management with AI For Dummies*. Hoboken, NJ: Wiley, 255–262.

Stokel-Walker C. 2022. AI bot ChatGPT writes smart essays—should professors worry? Nature. *Available at https://www.nature.com/articles/d41586-022-04397-7* (accessed 24 April 2025).

Stokel-Walker C. 2023. AI chatbots are coming to search engines—can you trust the results? Nature. *Available at https://www.nature.com/articles/d41586-023-00423-4* (accessed 24 April 2025).

Suppadungsuk S, Thongprayoon C, Miao J, Krisanapan P, Qureshi F, Kashani K, Cheungpasitporn W. 2023. Exploring the potential of chatbots in critical care nephrology. *Medicines* **10**:58 DOI 10.3390/medicines10100058.

Talyshinskii A, Naik N, Hameed BMZ, Juliebø-Jones P, Somani BK. 2024. Potential of AI-driven chatbots in urology: revolutionizing patient care through artificial intelligence. *Current Urology Reports* **25(1)**:9–18 DOI 10.1007/s11934-023-01184-3.

Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. 2023. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): are we ready yet? *Nurse Education Today* **129(8)**:105917 DOI 10.1016/j.nedt.2023.105917.

Tangsrivimol JA, Darzidehkalani E, Virk HUH, Wang Z, Egger J, Wang M, Hacking S, Glicksberg BS, Strauss M, Krittanawong C. 2025. Benefits, limits, and risks of ChatGPT in medicine. *Frontiers in Artificial Intelligence* **8**:1518049 DOI 10.3389/frai.2025.1518049.

Tebenkov E, Prokhorov I. 2021. Machine learning algorithms for teaching AI chat bots. *Procedia Computer Science* **190(4)**:735–744 DOI 10.1016/j.procs.2021.06.086.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. 2023. Large language models in medicine. *Nature Medicine* **29(8)**:1930–1940 DOI 10.1038/s41591-023-02448-8.

Turing AM. 1950. Computing machinery and intelligence. *Mind* **59(236)**:433–460 DOI 10.1093/mind/LIX.236.433.

Vaishya R, Iyengar KP, Patralekh MK, Botchu R, Shirodkar K, Jain VK, Vaish A, Scarlat MM. 2024. Effectiveness of AI-powered Chatbots in responding to orthopaedic postgraduate exam questions-an observational study. *International Orthopaedics* **48(8)**:1963–1969 DOI 10.1007/s00264-024-06182-9.

Wang S, Scells H, Koopman B, Zuccon G. 2023. Can ChatGPT write a good boolean query for systematic review literature search? In: *SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1426–1436 DOI 10.1145/3539618.3591703.