Original Article

A Comparative Analysis of the Performance of Large Language Models and Human Respondents in Dermatology

Aravind Baskar Murthy[1], Vijayasankar Palaniappan[2], Suganya Radhakrishnan[2], Sathish Rajaa[3], Kaliaperumal Karthikeyan[2]

[1]Department of Dermatology, Venereology and Leprosy, SRM Medical College Hospital and Research Center, Chengalpet, Tamil Nadu, [2]Department of Dermatology, Venereology and Leprosy, Sri Manakula Vinayagar Medical College and Hospital, Pondicherry, [3]Department of Community Medicine, ESIC Medical College and Hospital, Chennai, Tamil Nadu, India

Abstract

Background: With the growing interest in generative artificial intelligence (AI), the scientific community is witnessing the vast utility of large language models (LLMs) with chat interfaces such as ChatGPT and Microsoft Bing Chat in the medical field and research. This study aimed to investigate the accuracy of ChatGPT and Microsoft Bing Chat to answer questions on Dermatology, Venereology, and Leprosy, the frequency of artificial hallucinations, and to compare their performance with human respondents. **Aim and Objectives:** The primary objective of the study was to compare the knowledge and interpretation abilities of LLMs (ChatGPT v3.5 and Microsoft Bing Chat) with human respondents (12 final-year postgraduates) and the secondary objective was to assess the incidence of artificial hallucinations with 60 questions prepared by the authors, including multiple choice questions (MCQs), fill-in-the-blanks and scenario-based questions. **Materials and Methods:** The authors accessed two commercially available large language models (LLMs) with chat interfaces namely ChatGPT version 3.5 (OpenAI; San Francisco, CA) and Microsoft Bing Chat from August 10th to August 23rd, 2023. **Results:** In our testing set of 60 questions, Bing Chat outperformed ChatGPT and human respondents with a mean correct response score of 46.9 ± 0.7. The mean correct responses by ChatGPT and human respondents were 35.9 ± 0.5 and 25.8 ± 11.0, respectively. The overall accuracy of human respondents, ChatGPT and Bing Chat was observed to be 43%, 59.8%, and 78.2%, respectively. Of the MCQs, fill-in-the-blanks, and scenario-based questions, Bing Chat had the highest accuracy in all types of questions with statistical significance ($P < 0.001$ by ANOVA test). Topic-wise assessment of the performance of LLMs showed that Bing Chat performed better in all topics except vascular disorders, inflammatory disorders, and leprosy. Bing Chat performed better in answering easy and medium-difficulty questions with accuracies of 85.7% and 78%, respectively. In comparison, ChatGPT performed well on hard questions with an accuracy of 55% with statistical significance ($P < 0.001$ by ANOVA test). The mean number of questions answered by the human respondents among the 10 questions with multiple correct responses was 3 ± 1.4. The accuracy of LLMs in answering questions with multiple correct responses was assessed by employing two prompts. ChatGPT and Bing Chat could answer 3.1 ± 0.3 and 4 ± 0 questions respectively without prompting. On evaluating the ability of logical reasoning by the LLMs, it was found that ChatGPT gave logical reasoning in 47 ± 0.4 questions and Bing Chat in 53.9 ± 0.5 questions, irrespective of the correctness of the responses. ChatGPT exhibited artificial hallucination in 4 questions, even with 12 repeated inputs, which was not observed in Bing chat. **Limitations:** Variability in respondent accuracy, a small question set, and exclusion of newer AI models and image-based assessments. **Conclusion:** This study showed an overall better performance of LLMs compared to human respondents. However, the LLMs were less accurate than respondents in topics like inflammatory disorders and leprosy. Proper regulations concerning the use of LLMs are the need of the hour to avoid potential misuse.

Keywords: *Artificial Intelligence, ChatGPT, Bing Chat, Large Language Models, LLM*

Introduction

Generative pre-trained transformer (GPT) is a 175-billion-parameter natural language processing model that applies algorithms primed on a wide amount of data to provide output in a conversation style.[1–3] With the growing interest in generative artificial intelligence (AI), the scientific community is witnessing the vast utility of large language models (LLMs) with chat interfaces such as ChatGPT and Microsoft Bing Chat in various aspects of the medical field and research.[4] With the accuracy and integrity of these LLMs being tested in

How to cite this article: Murthy AB, Palaniappan V, Radhakrishnan S, Rajaa S, Karthikeyan K. A comparative analysis of the performance of large language models and human respondents in dermatology. Indian Dermatol Online J 2025;16:241-7.

Received: 10-Mar-2024. **Revised:** 14-Aug-2024. **Accepted:** 23-Aug-2024. **Published:** 27-Feb-2025.

Address for correspondence:
Dr. Vijayasankar Palaniappan, Department of Dermatology, Venereology and Leprosy, Sri Manakula Vinayagar Medical College and Hospital, Pondicherry, India.
E-mail: vijayasankar palaniappan@gmail.com

Access this article online

Website: https://journals.lww.com/idoj

DOI: 10.4103/idoj.idoj_221_24

Quick Response Code:

various aspects of the medical field, this study aimed to investigate the accuracy of ChatGPT and Microsoft Bing Chat to answer questions on Dermatology, Venereology, and Leprosy (DVL), the frequency of artificial hallucinations, and compare their performance with human respondents.

Methodology

The authors accessed two commercially available LLMs with chat interfaces, namely ChatGPT version 3.5 (OpenAI; San Francisco, CA) and Microsoft Bing Chat from 10 August to 23 August 2023. ChatGPT version 4.0 was not included in the study as sources suggest that Microsoft Bing Chat uses OpenAI's ChatGPT- v 4.0 as its LLM.[5,6]

The primary objective of the study was to compare the knowledge and interpretation abilities of LLMs (ChatGPT v3.5 and Microsoft Bing Chat) with human respondents (12 final-year postgraduates) by administering 60 questions prepared by the faculties of the DVL department in a tertiary care hospital. The secondary objective was to assess the incidence of artificial hallucinations.

Step 1: Selection of topics and blueprint of questions

Initially, an intradepartmental meeting was conducted with the faculties and the head of the department to devise a blueprint for the study. Based on the suggestions, the questions were chosen proportionately from all the possible topics of DVL. The questions were taken in such a way that the various formats of answering questions such as multiple choice questions (MCQs-35 questions), fill-in-the-blanks (16 questions), and real-time case scenarios (9 questions) were assessed among 60 questions. Out of the 35 MCQs, 10 questions had more than one correct response. Steps were taken to ensure that the questions were not directly taken from books or any online source that would favor the AI. To categorize the difficulty of questions, the authors individually scored the difficulty level of each question between 1 and 10. Questions were labeled as easy if the mean score was less than 3.9, as medium for scores between 4 and 6.9, and hard for scores 7 and above. In the end, easy, medium, and hard questions corresponded to 35, 15, and 10 questions, respectively. We avoided image - based questions because ChatGPT version 3.5 accepts only text input. Because ChatGPT version 3.5 is updated only till 2021, questions on any updates after 2021 were avoided to avoid bias in comparing the chat interfaces.[4]

Step 2: Conduction of the quiz among postgraduates

As a preliminary step, the postgraduates were assessed for 60 min. Apart from the total number of correct answers, the correct responses individually in MCQs, fill-in-the-blank type, and real-time scenarios, topic-wise correct responses were also assessed and mean scores were recorded.

Step 3: Uploading the questions in ChatGPT and Microsoft Bing Chat

To assess the performance and accuracy of AI in answering various types of questions in DVL, the same 60 questions were manually uploaded serially in two autoregressive language models (ALMs) namely ChatGPT version 3.5 and Microsoft Bing Chat. Screenshots of every response by the two models were also taken [Supplementary Files 1, http://links.lww.com/IDOJ/A7 and 2, http://links.lww.com/IDOJ/A8]. To generate mean and standard deviation (SD) for comparison with the results of the postgraduates, the questions were uploaded 12 times to assess consistency and generate mean and SD.

Step 4: Analysis of responses

The total number of correct and incorrect responses made by the postgraduates and LLMs was noted. In addition, the accuracy in different types of questions (MCQs, fill-in-the-blanks, and scenarios) and topic-wise responses were recorded. Evidence of artificial hallucination was documented.

Results

In our testing set of 60 questions, Bing Chat outperformed ChatGPT and human respondents with a mean correct response score of 46.9 ± 0.7. The mean correct responses by ChatGPT and human respondents were 35.9 ± 0.5 and 25.8 ± 11.0, respectively. The overall accuracy of human respondents, ChatGPT, and Bing Chat was observed to be 43%, 59.8%, and 78.2%, respectively. The accuracy of human respondents and LLMs in MCQs, fill-in-the-blanks, and scenario-based questions are tabulated in Table 1, with Bing Chat having the highest accuracy in all types of questions with statistical significance ($P < 0.001$ by analysis of variance [ANOVA] test).

Table 2 describes the performance of the LLMs and human respondents in the different topics of DVL. We noted that Bing Chat outperformed ChatGPT and human respondents in infections (85%), pilosebaceous disorders (84%), sexually transmitted diseases (80%), aesthetic dermatology (77.5%), therapeutics (76.7%), basic sciences (73.3%) and cutaneous adverse drug reactions (55%) with statistical significance (P value < 0.05). ChatGPT had the best accuracy in vascular disorders, whereas human respondents had better accuracy in inflammatory disorders and leprosy and the differences were found to be statistically significant ($P < 0.05$).

Bing Chat responded better than ChatGPT and human respondents in answering easy (30 ± 0.0) and medium - difficulty questions (11.7 ± 0.5) with accuracies of 85.7% and 78% respectively, whereas ChatGPT scored better than Bing Chat and human respondents in solving hard questions with an accuracy of 55%. This difference was found to be statistically significant with $P < 0.001$ by ANOVA test [Table 3].

The mean number of questions answered correctly by the human respondents among the 10 MCQs with multiple correct responses was 3 ± 1.4. To assess the accuracy of LLMs in answering questions with multiple correct responses, two prompts were employed. ChatGPT answered a mean of 3.1 ± 0.3 questions correctly without prompting, an additional 3.0 ± 0.4 questions with one prompt, and a further 2.9 ± 0.5 questions with two prompts, leaving 1.0 ± 0.7 questions unanswered even after two prompts. Bing Chat answered a mean of 4 ± 0 questions without prompting and 6 ± 0 questions could not be answered even after two prompts [Table 4]. Figures 1 and 2 summarize the distribution of responses by LLMs and human respondents with respect to the question type and difficulty level. On assessing the ability of logical reasoning by the LLMs, it was found that ChatGPT gave logical reasoning in 47 ± 0.4 questions and Bing Chat in 53.9 ± 0.5 questions, irrespective of the correctness of the responses.

The study assessed the frequency of artificial hallucinations among the LLMs. ChatGPT exhibited artificial hallucination in four questions, even with 12 repeated inputs, which was not observed in Bing Chat. The questions and the responses with artificial hallucinations by ChatGPT are shown in Figure 3. Table 5 summarizes the studies performed in this domain along with their results.[1,4,7–14]

Discussion

ChatGPT and Microsoft Bing Chat are autoregressive language models (ALMs), the latest among a class of LLMs, based on prompt engineering to encourage dialogic output.[1] The former ALM (ChatGPT) was introduced as a sibling model to InstructGPT in November 2022 as version 3.5.[2,3,15] Later ChatGPT version 4.0 and Microsoft Bing Chat were launched after fine-tuning the limitations of the older version. Microsoft Bing Chat has the additional advantage of actively surfing the internet and providing

Table 1: Comparative analysis of accuracies of large language models and human respondents based on question types

Type of questions	Total number of questions	Total number of correct by human respondents (mean ± SD)	Total correct by Chat GPT	Total correct by Bing Chat	*P**
MCQ	35	17±6.3	21.1±0.3	24.7±0.5	<0.001
Fill ups	16	5.3±3.6	9.7±0.5	15±0.0	<0.001
Scenario	9	3.6±2.2	5.2±0.7	7.1±0.3	<0.001
Total	60	25.8±11.0	35.9±0.5	46.9±0.7	<0.001

*Significant by ANOVA (comparing the mean of three different groups). MCQ - Multiple choice questions

Table 2: Topic-wise comparison of the performance of large language models and human respondents

Topics	Number of questions	Total number of correct by human respondents - Mean ± SD (%)	Total correct by ChatGPT- Mean ± SD (%)	Total correct by Bing Chat-Mean ± SD (%)	*P**
Basic sciences	3	1.7 ± 0.2 (56.7%)	2.1 ± 0.2 (70%)	2.2 ± 0.3 (73.3%)	<0.001
Infections	4	2.3 ± 0.4 (57.5%)	2.1 ± 0.2 (52.5%)	3.4 ± 0.1 (85%)	0.024
Inflammatory diseases	5	3.1 ± 0.5 (62%)	0.4 ± 0.2 (12.9%)	2.2 ± 0.7 (44%)	<0.001
Connective tissue disorders	4	2.0 ± 0.2 (50%)	3.4 ± 0.2 (85%)	3.2 ± 0.3 (80%)	0.056
Genodermatoses	6	3.1 ± 0.7 (51.7%)	5.2 ± 0.3 (86.7%)	5.2 ± 0.3 (86.7%)	0.047
Metabolic disorders	5	3.0 ± 0.3 (60%)	4.3 ± 0.2 (86%)	4.4 ± 0.3 (88%)	0.061
Vascular diseases	4	1.9 ± 0.4 (47.5%)	3.5 ± 0.2 (87.5%)	3.4 ± 0.4 (85%)	0.011
Pilosebaceous disorders	5	2.5 ± 0.2 (50%)	3.2 ± 0.3 (64%)	4.2 ± 0.3 (84%)	<0.001
Cutaneous adverse drug reactions	4	1 ± 0.3 (25%)	2.1 ± 0.2 (52.5%)	2.2 ± 0.3 (55%)	0.010
Benign and malignant tumours	4	2.1 ± 0.3 (52.5%)	2.2 ± 0.2 (55%)	3.2 ± 0.2 (80%)	0.058
Therapeutics	3	1.5 ± 0.2 (50%)	1.1 ± 0.3 (36.7%)	2.3 ± 0.2 (76.7%)	<0.001
Leprosy	5	3.4 ± 0.2 (68%)	1.1 ± 1.1 (22%)	3.1 ± 0.2 (62%)	<0.001
Sexually transmitted diseases	4	2.1 ± 0.4 (52.5%)	3.1 ± 0.2 (77.5%)	3.2 ± 0.2 (80%)	0.023
Aesthetic dermatology	4	1.4 ± 0.3 (35%)	2.1 ± 0.5 (52.5%)	3.1 ± 0.3 (77.5%)	<0.001
Total	60	25.8±11.0	35.9±0.5	46.9±0.7	<0.001

*Significant by ANOVA (comparing the mean of three different groups)

Table 3: Difficulty-wise accuracy of large language models and human respondents

Difficulty	Human respondents	ChatGPT	Bing Chat	*P**
Easy (35)	16.4±6.8 (46.8%)	22±0.0 (62.8%)	30±0.0 (85.7%)	<0.001
Medium (15)	6.7±3.1 (44.7%)	8.4±0.5 (56%)	11.7±0.5 (78%)	<0.001
Hard (10)	2.8±1.9 (28%)	5.5±0.5 (55%)	5.1±0.3 (51%)	<0.001
Total (60)	25.8±11.0	35.9±0.5	46.9±0.7	<0.001

*Significant by ANOVA (comparing the mean of three different groups)

Table 4: Comparison of accuracies of large language models and human respondents to questions with multiple correct responses (*n*=10)

	Answered correctly without prompting	One prompt needed	Two prompts needed	No correct response even after two prompts
Human respondent	3	N/A	N/A	N/A
ChatGPT	3	3	3	1
Bing Chat	4	0	0	6

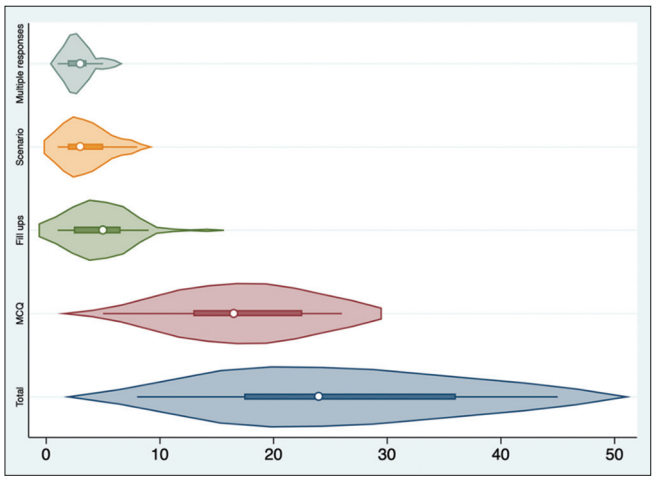

Figure 1: Distribution of responses by LLMs and human respondents with respect to question type

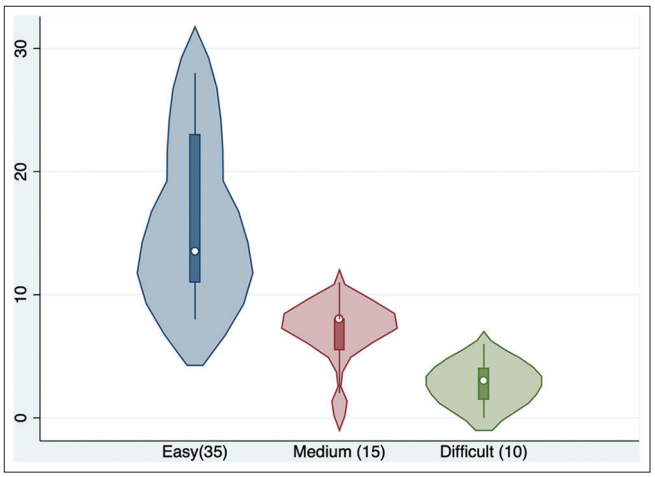

Figure 2: Distribution of responses by LLMs and human respondents with respect to difficulty level

relevant sources for the output.[13] The exceptional speed, intelligibility, and anthropomorphic nature of outputs of GPT chatbots have attracted the attention of the medical fraternity. The architecture of GPT employs a neural network to process the natural language, thereby generating input-based responses.[16]

Recently the performance of LLMs such as ChatGPT and Bing Chat in various licensing examinations to assess their primary competency of medical knowledge is on the growing trend.[1,12–14] Studies on the performance of ChatGPT on USMLE exam questions found that the AI was able to or almost reach the passing threshold with 60% accuracy which has increased from 36.7% over the past 3 years.[1,14]

In this study, Bing Chat was found to have the highest accuracy (78.2%), outperforming ChatGPT-3.5 (59.8%) and human respondents (43%), which was similar to other studies.[4,13] Though Bing Chat had the highest accuracy in all subtopics of DVL, ChatGPT, and human respondents scored well in a few topics. Similarly, a study on LLM's performance in ophthalmology board-style questions found human respondents' average to be better than LLMs in a few topics and similar human domination in another Korean study.[4,12]

Regardless of the type of questions, Bing Chat performed better than ChatGPT-v3.5 and human respondents in MCQs, scenarios, and fill-in-the-blanks questions. We observed that the LLMs performed only on par with the human respondents in questions with multiple correct responses. Similarly, Hoch CC *et al.*[8] and Huh *et al.*,[12] also found that MCQs with multiple correct responses were a great hurdle for LLMs, probably attributed to their underlying operational principles. It could be understood that the LLMs are designed to obtain the most plausible response rather than the evaluation of each option independently.

It was observed that the LLMs had better accuracy with fill-in-the-blank questions than MCQs and scenario-based questions, whereas human respondents performed well with the MCQs. Because MCQs contain options that vary in meaning and context, it could struggle to identify subtle differences in the options provided, requiring additional training for better performance.

Bing Chat showed the highest accuracy in answering easy and medium-difficulty questions but ChatGPT scored well in hard questions. This was concordant with the findings observed in the study by Cai *et al.*,[4] in which the accuracy of ChatGPT v3.5 and Bing Chat in answering easy questions was 70.3% and 82.9%, respectively, whereas the accuracy for hard questions was 39.4% and 34.4%. In this study, logical explanation was given for 78.3% and 89.8% of the questions by ChatGPT and Bing Chat, respectively. Another finding observed was that the LLMs were able to provide logical explanations in all the questions (100%) with multiple correct responses even during the prompts. A study by Gilson *et al.*[1] found that a logical explanation

Table 5: Summary of studies done on assessing the performance of large language models in various medical specialties

Study Title	Author	Results
Exploring the Potential and Limitations of Chat Generative Pre-trained Transformer (ChatGPT) in Generating Board-Style Dermatology Questions: A Qualitative Analysis	Ayub I *et al.* (2023)[7]	• Out of 40 questions generated, only 16 (40%) were deemed accurate and appropriate for ABD-AE study preparation • The remaining questions exhibited limitations, including low complexity, lack of clarity, and inaccuracies.
How does ChatGPT perform on the United States Medical Licensing examination? The implications of Large language models for medical education and knowledge assessment (2023)	Gilson A *et al.* (2023)[1]	• The study used 2 pairs of data consisting of 100 and 120 questions respectively • Among *AMBOSS-Step1, AMBOSS-Step2, NBME-Free-Step1, and NBME-Free-Step2*, ChatGPT achieved accuracies of 44% (44/100), 42% (42/100), 64.4% (56/87), and 57.8% (59/102), respectively • Logical justification for ChatGPT's answer selection was present in 100% of the outputs of the *NBME* data sets.
Performance of Generative Large Language Models on Ophthalmology Board–Style Questions	Cai LZ (2023)[4]	• Of the total 250 questions assessed, human respondents had an average accuracy of 72.2%. ChatGPT-3.5 scored the lowest (58.8%), whereas ChatGPT-4.0 (71.6%) and Bing Chat (71.2%) performed comparably. • ChatGPT-4.0 and Bing Chat struggled with image interpretation when compared with single-step reasoning questions. • ChatGPT-3.5 had the highest rate of hallucinations and nonlogical reasoning (42.4%), followed by ChatGPT-4.0 (18.0%) and Bing Chat (25.6%)
ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single choice and multiple choice board certification preparation questions	Hoch HH *et al.* (2023)[8]	• ChatGPT answered 57% (n=1,475) of the total number of questions (n=2,576) • Single-choice questions were associated with a significantly higher rate (P<0.001) of correct responses compared to multiple-choice questions
Comparing Meta-Analyses with ChatGPT in the Evaluation of the Effectiveness and Tolerance of Systemic Therapies in Moderate-to-Severe Plaque Psoriasis	Lam Hoai, X.-L *et al.* (2023)[9]	• The study analyzed whether ChatGPT was able to summarize information in a useful fashion for providers and patients in a way that matches up with the results of 3 meta-analyses (MAs) and 13 network meta-analyses (NMAs) • The reproducibility between the results of ChatGPT and the MAs/NMAs was random regarding drug safety. • Regarding efficacy, ChatGPT reached the same conclusion in 5 out of the 16 studies (four out of four studies when three molecules were compared), gave acceptable answers in 7 out of 16 studies, and was inconclusive in 4 out of 16 studies
Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model	Johnson D *et al.* (2023)[10]	• Across all questions (n=284), the median accuracy score was 5.5 (between almost completely and completely correct) with a mean score of 4.8 (between mostly and almost completely correct) (on a Likert scale between 1 and 6, with 6 being completely correct) • For questions rated easy, medium, and hard, median accuracy scores were 6, 5.5, and 5 • Accuracy scores for binary and descriptive questions were similar
Performance of the Large Language Model ChatGPT on the National Nurse Examinations in Japan: Evaluation Study	Taira K *et al.* (2023)[11]	• Questions on the Japanese National Nurse Examinations from 2019 to 2023 (240 questions each year) were used for assessing the performance of ChatGPT 3.5. • The 5-year average percentage of correct answers for ChatGPT was 75.1% (SD 3%) for basic knowledge questions and 64.5% (SD 5%) for general questions. • ChatGPT met the passing criteria for the 2019 Japanese National Nurse Examination and was close to passing the 2020-2023 examinations, with only a few more correct answers required to pass
Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study	Huh S *et al.* (2023)[12]	• Out of 79 questions, ChatGPT's performance (60.8% accuracy) was lower than that of the medical students (90.8%) • There was a significant relationship between acceptable explanations and correct answers (P=0.0000)

Contd...

Table 5: Contd...

Study Title	Author	Results
Artificial intelligence in medical education: Comparative analysis of ChatGPT, Bing and medical students in Germany	Roos J *et al.* (2023)[13]	• On assessing 630 questions from the spring and fall German Medical State Examinations of 2022, GPT-4 achieved the highest overall performance, correctly answering 88.1% of questions, closely followed by Bing (86.0%) and GPT-3.5-Turbo (65.7%) • The students had an average correct answer rate of 74.6%. Both GPT-4 and Bing significantly outperformed the students in both examinations. • When media questions were excluded, Bing achieved the highest performance of 90.7%, closely followed by GPT-4 (90.4%), whereas GPT-3.5-Turbo lagged (68.2%).
Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models	Kung TH (2023)[14]	• ChatGPT performed at or near the passing threshold of 60% accuracy and displayed comprehensible reasoning and valid clinical insights

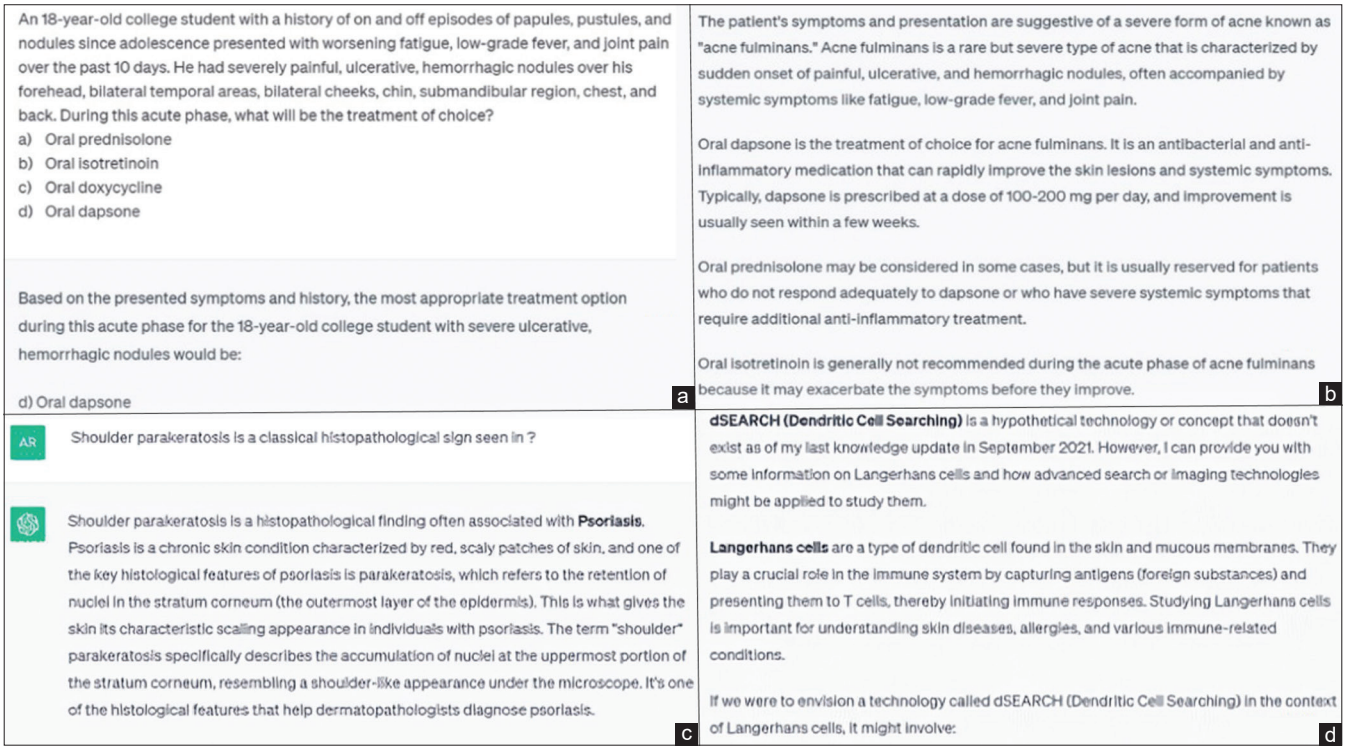

Figure 3: Artificial hallucinations observed in the responses of ChatGPT

was provided for every question uploaded, irrespective of the correctness of the response.

Artificial hallucination is a phenomenon in LLMs in which seemingly realistic sensory experiences are generated that do not correspond to the real-world input.[3] This study found that the frequency of artificial hallucinations was 6.7% (4 questions) in ChatGPT v3.5, whereas Bing Chat did not hallucinate for any question. In a study analyzing the performance of generative LLMs on ophthalmology board-style questions, it was found that ChatGPT v3.5 had the highest frequency of artificial hallucinations and non-logical reasoning (42.4%), whereas ChatGPT version 4.0 and Bing Chat reported lower frequencies of 18% and 25.6%, respectively.[4]

While the applications of these ALMs in the medical and research fields are engrossing, concerns have been raised about the potential risks of GPTs in providing inaccurate data, illogical reasoning, artificial hallucinations, poor accuracy in questions with multiple correct responses, and cyberattacks with the spread of misinformation.[17,18] The utility of ChatGPT in academic writing as assignments among medical undergraduates and scientific paper writing among researchers threatens to weaken thinking and writing abilities.[19]

Limitations

The limitation of the study was the use of mean human respondent accuracy with the performance of LLMs. Each

student had different learning capabilities and problem-solving skills. Hence, even when a few students scored better than the LLMs, the mean score fell below the LLMs. The study also used fewer questions and human respondents for comparison with the LLMs. Exclusion of image-based assessment was another limitation because ChatGPT v3.5 did not support image inputs. The inclusion of image inputs in future studies can further compare the visual interpretation skills of LLMs and human respondents and suggest areas for improvement. Because questions with updates after 2021 were avoided, the knowledge of current trends in healthcare could not be assessed in this study. Moreover, newer LLMs such as Gemini, launched after the completion of the study, could not be included for analysis and Bing Chat has been replaced by Copilot, as a result of the rapid evolution of AI.

Conclusion

This study showed an overall better performance of LLMs compared to human respondents. However, the LLMs were less accurate than human respondents in topics such as inflammatory disorders and leprosy. Further studies exploring the performance of LLMs in various medical specialties will provide insights into their potential and limitations, thereby defining their role in medical education and research. The artificial hallucinations observed with LLMs in a few questions warrant continuous monitoring by experts to avoid misconceptions, especially when used in health care. Proper regulations concerning the use of LLMs are the need of the hour to avoid potential misuse. The further growth of AI in medical research, medical education, and healthcare is quite enticing.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, *et al*. How does ChatGPT Perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312.
2. Scott K. Microsoft teams up with OpenAI to exclusively license GPT-3 language model. The Official Microsoft Blog. 2020. Available from: https://blogs.microsoft.com/blog/2020 / 09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/. [Last accessed on 2024 Feb 20].
3. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus 2023;15:e35179.
4. Cai LZ, Shaheen A, Jin A, Fukui R, Yi JS, Yannuzzi N, *et al*. Performance of generative large language models on ophthalmology board-style questions. Am J Ophthalmol 2023;254:141–9.
5. Shields W. Microsoft Bing Gets a Brain Upgrade with OpenAI's GPT-4 A.I. LinkedIn. Available from: https://www.linkedin.com/pulse/microsoft-bing-gets-brain-upgrade-openais-gpt-4-ai-walter-shields/. [Last accessed on 2024 Feb 20].
6. Jones L. Microsoft Bing Chat to Get GPT-4 Upgrade, Adding Video and Audio Responses. WinBuzzer. 2023. Available from: https://winbuzzer.com/2023 / 03/10/microsoft-bing-chat-to-get-gpt-4-upgrade-adding-video-and-audio-responses-xcxwbn/. [Last accessed on 2024 Feb 20].
7. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: A qualitative analysis. Cureus 2023;15:e43717.
8. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, *et al*. ChatGPT's quiz skills in different otolaryngology subspecialties: An analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023;280:4271-8.
9. Lam Hoai XL, Simonart T. Comparing meta-analyses with ChatGPT in the evaluation of the effectiveness and tolerance of systemic therapies in moderate-to-severe plaque psoriasis. J Clin Med 2023;12:5410.
10. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, *et al*. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT Model. Res Sq 2023;rs.3.rs-2566942.
11. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: Evaluation study. JMIR Nurs 2023;6:e47305.
12. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. J Educ Eval Health Prof 2023;20:1.
13. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial intelligence in medical education: Comparative analysis of ChatGPT, Bing, and medical students in Germany. JMIR Med Educ 2023;9:e46482.
14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al*. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198.
15. Introducing ChatGPT. Available from: https://openai.com/blog/chatgpt. [Last accessed on 2024 Feb 20].
16. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, *et al*. Language models are few-shot learners. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. p. 1877–901.
17. Deng J, Lin Y. The Benefits and challenges of ChatGPT: An overview. Front Comput Intell Syst 2022;2:81–3.
18. Beutel G, Geerits E, Kielstein JT. Artificial hallucination: GPT on LSD? Crit Care Lond Engl 2023;27:148.
19. Chatterjee J, Dethlefs N. This new conversational AI model can be your friend, philosopher, and guide. and even your worst enemy. Patterns 2023;4:100676.