Evaluating the Ability of Chatbots to Answer Entrance Exam Questions for Postgraduate Studies in Medical Laboratory Sciences in Iran

Farhad AREFINIA[1,2], Azamossadat HOSSEINI [1,*], Farkhondeh ASADI [1,*], Versa OMRANI-NAVA [2] and Raham NILOOFARI[3]

[1] Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Darband Str., no. 41, 1971653313, Tehran, Iran.
[2] Department of Paramedical Sciences, Amol School of Paramedicine sciences, Mazandaran University of Medical Sciences, Valiasr Str., 4815733971, Sari, Iran.
[3] Student Research Committee, School of Paramedical Sciences, Mazandaran University of Medical Sciences, Talebamoli Str., no. 22.1, 4615861468, Amol, Iran.
E-mails: farhad.arefinia@mazums.ac.ir; *souhosseini@sbmu.ac.ir; *asadifar@sbmu.ac.ir; versa.omraninava@mazums.ac.ir; raham.niloofari@mazums.acom

* Author to whom correspondence should be addressed;

Received: 18 June 2024/Accepted: 26 September 2024/ Published online: 29 September 2024

Abstract
As educational technology advances, the integration of Artificial Intelligence (AI)-driven chatbots in academic contexts becomes increasingly relevant. This study explored the performance of three advanced chatbots—ChatGPT 3.5, Claude, and Google Bard—in responding to entrance exam questions for Master's and PhD. programs in Medical Laboratory Sciences in Iran in 2023. Multiple-choice questions from entrance exams in Medical Laboratory Sciences Master's and PhD. programs held in 2023 were presented to ChatGPT 3.5, Claude, and Google Bard, and their responses were evaluated. The three chatbots—ChatGPT 3.5, Claude, and Google Bard—exhibited an overall accuracy of 38%, 42%, and 37%, respectively, showcasing a comparable baseline proficiency in addressing a variety of questions. Subject-specific analysis highlighted their strengths and weaknesses in different scientific domains. Our study shows that while the evaluated chatbots showed some ability in answering medical laboratory science questions, their performance remains insufficient for success in postgraduate entrance exams.

Keywords: Chatbots; Medical Laboratory Sciences; Education, Multiple-Choice Questions Exams; Comparative Analysis

Introduction

Over the past decades, Iran is facing a significant scientific growth, an increase in the number of universities and applicants for postgraduate education [1]. Medical Laboratory science, included in the group of paramedical disciplines, is a medical science subset dealing with disease diagnosis, providing critical data for accurate diagnosis, treatment planning, and disease monitoring. In addition to clinical or research activities, graduates of these fields can participate in the entrance exam for master's degrees and then specialized doctorates in various fields [2]. One of the inseparable parts of the teaching process is the evaluation of the level of learning and academic progress, which can determine the achievement of the educational goals. In this regard, multiple-choice questions are a common evaluation method in university systems and education applicants around the world. They are easy to correct and show good reliability [3].

Integration of Artificial Intelligence (AI) with medical education holds immense potential for transforming the teaching and learning methodologies in the field of medical sciences [4]. Currently, AI is constantly getting better and is seen as a technology that changes everything [5]. Researchers even coin the term 'Fourth Industrial Revolution' to capture the transformative impacts AI is exerting [6].

AI is continuously developed and is recognized as a strong reality now. Rapid advances in AI have led to the development of large language models (LLM) that have an unprecedented ability to understand and perform natural language processing (NLP) tasks [7]. ChatGPT was the first model unleashed by OpenAI on November 30, 2022, and soon became popular among academic users. In academia, ChatGPT supports research and learning through Promote creativity and innovation, improve data analysis and interpretation, provide additional resources and Support research in an emerging technology but it also raises concerns about misinformation, academic integrity, and over-reliance on AI. Ethical issues regarding authorship, originality, and fairness highlight the need for clear guidelines on AI use in academic settings [8].

Google Bard, developed and introduced by Google in 2023 [9], represents another notable advancement in the field of AI-driven chatbots. Likewise, Claude, an experiential chatbot created by Anthropic that was introduced in 2023 is another significant indicator of AI conversational agents [10]. While there are inconsistencies regarding their performance, the mentioned bots show promising performance in educational purposes [11,12].

A variety of utilities are mentioned for Chatbots including helping with medical reports and suggesting treatment options and abilities to perform different stages of research projects [13]. In Medical Education particularly in paramedical courses; AI chatbots can prove to be highly beneficial in offering customized learning, sample questions, immediate feedback, individual learning paths, and supplementary material. These tools can further improve student's knowledge, retention and prepare them for tests leading to better results in their paramedical examinations. To recap, although Chatbots do not exclude traditional means of knowledge delivery, they can assist in making the process less random and unproductive [14,15]. However, LLMs are not without limitations and sometimes show poor performance in some medical categories especially conceptual topics that require interpretation or more complex questions such as board exams [16].

Considering the importance LLMs and chatbots in medical education and related fields including medical laboratory sciences, this study aimed to evaluate and compare the ability of free-access version of three chatbots including ChatGPT 3.5, Google Bard, Claude and establish a baseline for comparison with newer models of these Chatbots through answering medical laboratory Science Master's and Ph.D. entrance exams held by the Ministry of Medical Education of Iran in 2023.

Materials and Methods

This was a cross-sectional comparative study and was done between December 2023 to January 2024. the entrance exam multiple-choice questions of Medical Laboratory Science Master's and PhD. programs in 2023 were collected from the official website of the Iranian National Organization for Educational Testing. The exam questions consisted of 1160 Persian four-option multiple-choice questions related to 13 courses, with 51 questions excluded from the analysis due to the presence of formulas and figures.

The questions were entered one by one with their options into the dialogue box of the following chatbots:

- ChatGPT (version 3.5), an AI system developed by OpenAI.
- Google Bard, developed and introduced by Google in 2023.
- Claude, an experimental chatbot by Anthropic launched in 2023.

Each question was presented to the chatbots using the following standardized prompt: 'Please answer the following multiple-choice question from a Medical Laboratory Sciences exam. Provide only the letter corresponding to the correct answer (A, B, C, or D). " No additional context or instructions were given.

The chatbots' responses were individually collected in an Excel sheet. Two blinded reviewers evaluated the accuracy of the responses by comparing them to the key answers and assigned a binary score (0 for incorrect, 1 for correct). According to vice-chancellor of medical education in Iran, the passing score for the entrance exam for the Medical Laboratory Sciences Master's degree and the PhD. program is determined to be at least 30% correct answers after negative marking is calculated. The raw scores were converted into percentages by dividing the number of correct responses by the total number of questions, and no adjustments for negative marking were applied.

Statistics. Descriptive statistics including percentages of correct responses for each chatbot were calculated using Excel. The chatbots' overall performance in answering the questions was evaluated. Chi-Squared test was applied to compare correct answer rate between bots considering P<0.05 as statistically significant. Data analysis was performed by GraphPad Prism software.

Results

A total of 1109 multiple-choice questions were presented to three chatbots: ChatGPT, Claude, and Google Bard. Table 1, summarizes the overall chatbot performance across all course subjects

Collectively, ChatGPT accurately responded to 38% of the questions. Claude exhibited a comparable overall accuracy of 42%. Google Bard demonstrated an overall accuracy of 37% across all questions.

Overall, the three chatbots showcased similar performances in responding to specialized exam questions related to medical laboratory sciences, with no significant discrepancies in their overall accuracy (p-value=0.07). Their proficiency ranged from approximately 30% to 45% across different subjects.

Table 1. Results of the ability of three chatbots to answer questions, split by courses

Course	**Number of Questions**	**ChatGPT Correct Answers (Accuracy %)**	**Claude Correct Answers (Accuracy %)**	**Google Bard Correct Answers (Accuracy %)**	**P-value**
Parasitology	85	26 (31)	36 (42)	32 (38)	0.27
Immunology	176	71 (40)	82 (47)	62 (35)	0.09
Bacteriology	76	34 (45)	21 (28)	29 (38)	0.08
Biochemistry	141	57 (40)	56 (40)	55 (39)	0.97
Protozoology	20	9 (45)	9 (45)	7 (35)	0.76
Hematology	92	35 (38)	39 (42)	33 (36)	0.65
Cellular and Molecular Biology	125	51 (40)	55 (44)	56 (45)	0.79
Human Genetics	98	39 (40)	50 (51)	33 (34)	0.04
Organic and General Chemistry	14	2 (14)	9 (64)	7 (50)	0.02
Blood transfusion science	33	12 (36)	10 (30)	12 (36)	0.83
Mycology	104	34 (33)	44 (42)	34 (33)	0.24
Microbiology	43	15 (35)	17 (40)	20 (47)	0.54
Virology	102	37 (36)	36 (35)	35 (34)	0.95
Total	**1109**	**422 (38)**	464 (**42**)	415 (**37**)	0.07

Table 2 breaks down results by PhD. and Master's questions. ChatGPT accuracy varied from 14% in organic and general chemistry to 50% in bacteriology among Master's level courses. Notably Claude demonstrated the highest accuracy (64%) in organic and general chemistry among all subjects and chatbots tested. but demonstrated a weaker performance in blood transfusion science with 12% accuracy among PhD level courses.

Google Bard achieved the highest accuracy of 50% in organic and general Chemistry and in Virology for Master's exams, while its lowest accuracy, 16%, was observed in Mycology questions for Master's exams. Claude showed better performance in Ph.D. questions (P=0.04).

Table 2. Results of the ability of three chatbots to answer on PhD. and master's questions

Level	Course	Number of Questions	ChatGPT Correct Answers (Accuracy %)	Claude Correct Answers (Accuracy %)	Google Bard Correct Answers (Accuracy %)	P-value
PhD.	Parasitology	65	20 (31)	30 (46)	27 (42)	0.18
	Immunology	135	55 (41)	65 (48)	49 (36)	0.13
	Bacteriology	56	24 (43)	16 (29)	24 (43)	0.19
	Biochemistry	89	38 (43)	36 (40)	34 (38)	0.82
	Protozoology	20	9 (45)	9 (45)	7 (35)	0.76
	Hematology	64	26 (41)	33 (52)	21 (33)	0.09
	Cellular and Molecular Biology	68	26 (38)	29 (43)	32 (47)	0 .58
	Human Genetics	76	29 (38)	40 (53)	25 (33)	0.03
	Blood transfusion science	33	12 (36)	10 (30)	12 (36)	0.83
	Mycology	85	28 (33)	36 (42)	31 (36)	0.43
	Virology	84	32 (38)	30 (36)	26 (31)	0.61
	Total of Ph.D. questions	**775**	**299 (39)**	**334 (43)**	**288 (37)**	**0.04**
master's	Parasitology	20	6 (30)	6 (30)	5 (25)	0.92
	Immunology	41	16 (39)	17 (41)	13 (32)	0.63
	Bacteriology	20	10 (50)	5 (25)	5 (25)	0.15
	Biochemistry	52	19 (37)	20 (38)	21 (40)	0.92
	Hematology	28	9 (32)	6 (21)	12 (43)	0.22
	Cellular and Molecular Biology	57	25 (44)	26 (46)	24 (42)	0.93
	Human Genetics	22	10 (45)	10 (45)	8 (36)	0.78
	Organic and General Chemistry	14	2 (14)	9 (64)	7 (50)	0.02
	Mycology	19	6 (32)	8 (42)	3 (16)	0.20
	Microbiology	43	15 (35)	17 (40)	20 (47)	0.54
	Virology	18	5 (28)	6 (33)	9 (50)	0.35
	Total of master's questions	**334**	**123 (37)**	**130 (39)**	**127 (38)**	**0.85**

Discussion

Overall Chatbot Performance

The present study showed that the 3 evaluated chatbots (ChatGPT, Claude and Bard) performed almost similarly to each other in answering the questions of PhD and Master's entrance exam of laboratory sciences in Iran, however, Claude achieved relatively better performance. ChatGPT, Claude, and Google Bard exhibited an overall accuracy of 38%, 42%, and 37%, respectively. Although Claude showed better overall performance, the remarkable alignment in performance suggests a comparable baseline proficiency in addressing the presentation of questions posed. It has been reported that Claude performs superior than GPT-4 and Bard in the category of facts data [17]. Also, in both fields of diagnosis and treatment decisions for head and neck squamous cell carcinoma, Claude achieved superior outcomes compared to ChatGPT 4.0 [18]. The observed P-value indicates the possibility of inherent differences that could become more apparent with an increased sample size or within particular subjects.

Comparisons Between Chatbots

Chatbots can not provide sources for their information, which is a big downside. Also, accuracy of information needs to be confirmed which is really important [19].

The results are disparate similar with Kataoka et al which investigated Japanese Medical Licensure Examination [20]. GPT-3.5 correctly answered 57.4% of Family and Community Medicine Progress multiple-choice questions as provided by Huang [21]. The obtained value was 46% for ophthalmology board certification questions [15]. In contrast, higher correct answer rate (81.3%) was observed when Iranian Medical Residency Examination were fed to ChatGPT [6]. Plevris et al. [22] reported that these chatbots performed acceptably on simple mathematical and logical problems, but showed decreased performance when faced with complex problems.

Strengths and Weaknesses of Each Chatbot

Our study also shows that, given the difficulty of master's and Ph.D. level questions, the overall performance of chatbots has not been acceptable. This is because in the tests, the final score after deducting the negative score must be at least 30%, and the values reported in the present study are raw scores. ChatGPT achieved 50% in Bacteriology among Master's questions. The variability in results among the subjects and tiers provides a good insight into strengths and weaknesses in these AI models. The high scores, such as ChatGPT with 50% at the Master's level in Bacteriology and Claude with 64% at the Master's level in Organic Chemistry, may relate to large and high-quality training data. In contrast, the low scores-in instances like that of Google Bard for 16% Master's-level Mycology-may indicate weaknesses in the knowledge base or challenges of the subject matter being specialty-specific.

Implications for Medical Education

The use of Persian language in the exam questions may have impacted the chatbots' performance. The AI models are primarily trained on English language data, which could lead to misinterpretation or inaccurate translations of specialized medical terminology. This limitation might have resulted in lower accuracy scores than if the questions were presented in English. Future studies could address this by comparing chatbot performance on equivalent sets of questions in multiple languages to quantify the impact of language on accuracy.

The multiple-choice format of the questions may have both helped and hindered the chatbots' performance. On the one hand, it could have made some questions easier by providing possible answers. On the other hand, it might have limited the chatbots' ability to demonstrate more nuanced understanding. Future research could compare chatbot performance on multiple-choice questions versus open-ended questions to better understand the impact of question format on AI accuracy.

Conclusions

Our study demonstrated that free-access version of three chatbots including ChatGPT, Google Bard, and Claude, which were leading at the time of conducting this study, have insufficient ability in answering specialized questions from entrance examinations for postgraduate studies in medical laboratory sciences in Iran and require expert oversight due to potential errors for medical student. Although their overall performance ranged from 37% to 42%, this level of accuracy is insufficient for passing the entrance examination because the mentioned exams calculate negative ranks on final score too. However, the field of conversational artificial intelligence is rapidly advancing, and newer models are likely to perform better on these examinations. The present study provides a baseline of the chatbots' abilities at the time of the study, which can be used for comparison with newer models. This study can serve as a highly useful benchmark for demonstrating future advancements in chatbots' ability to respond to questions in future research.

List of Abbreviations: AI: Artificial Intelligence; LLM: Large Language Models; NLP: Natural Language Processing; Ph.D.: Doctor of Philosophy

Author Contributions: FA, AH, and FA conceptualized and designed the study. FA and RN managed data gathering and analyses. FA and VO contributed to writing and critical evaluation. All authors reviewed and approved the final manuscript draft.

Funding: This research received no funding.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Azadbakht L, Haghighatdoost F, Esmaillzadeh A. Favorable outcomes of using critical appraisal technique beside lecturing method for teaching theoretical sections. Iranian Journal of Medical Education 2011;10:651-658.
2. Rafieemehr H, Rostami Moez M. Assessing of the degree of compliance of the undergraduate laboratory sciences curriculum with the job skills based on responsive education in Hamadan University of Medical Sciences in 1399. Horizons of Medical Education Development 2022;13:26-15.
3. Khajeali N, Aslami M, Araban M. Analysis of medical and dentistry basic sciences examinations: A case study. Payesh (Health Monitor) 2020;19:383-389.
4. Liu J, Liu S. The application of ChatGPT in medical education. PrePrint 2023. doi:10.35542/osf.io/wzc2h
5. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT-Reshaping medical education and clinical management. Pakistan Journal of Medical Sciences 2023;39:605. doi:10.12669/pjms.39.2.7653.
6. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M, Acharya UR, Abbasian Ardakani A. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. Informatics in Medicine Unlocked 2023;41:101314. doi:10.1016/j.imu.2023.101314.
7. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Sci Rep 2023;13:20512. doi:10.1038/s41598-023-46995-z.
8. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee S-S. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Frontiers in Artificial Intelligence 2023;6:1237704. doi:10.3389/frai.2023.1237704.
9. Labadze L, Grigolia M, Machaidze L. Role of AI chatbots in education: systematic literature review. International Journal of Educational Technology in Higher Education 2023;20:56. doi:10.1186/s41239-023-00426-1.
10. Garcia Valencia OA, Suppadungsuk S, Thongprayoon C, Miao J, Tangpanithandee S, Craici IM, Cheungpasitporn W. Ethical Implications of Chatbot Utilization in Nephrology. Journal of Personalized Medicine 2023;13:1363. doi: 10.3390/jpm13091363.
11. Sarbay İ, Bozdereli̇ Beri̇Kol G, ÖZturan İU, Gri̇Mes K. Comparison of Performances of Open Access Natural Language Processing Based Chatbot Applications in Triage Decisions. Kırıkkale Üniversitesi Tıp Fakültesi Dergisi 2023;25:482-521. doi:10.24938/kutfd.1369468.
12. Lozić E, Štular B. Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. Future Internet. 2023;15(10):336. doi:10.3390/fi15100336.
13. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: Current Status and Future Directions. Journal of Multidisciplinary Healthcare 2023:1513-1520.
14. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study. JMIR Med Educ 2023, 9:e50514. doi:10.2196/50514.
15. Mihalache A, Popovic MM, Muni RH. Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment. JAMA Ophthalmology 2023;141:589-597. doi:10.1001/jamaophthalmol.2023.1144.
16. Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. Diagnostic Pathology 2024;19:43. doi: 10.1186/s13000-024-01464-7.
17. Borji, Ali and Mohammadian, Mehrdad, Battle of the Wordsmiths: Comparing ChatGPT, GPT-4, Claude, and Bard (June 12, 2023). doi:10.2139/ssrn.4476855.

18. Schmidl B, Hütten T, Pigorsch S, Stögbauer F, Hoch CC, Hussain T, Wollenberg B, Wirth M. Assessing the use of the novel tool Claude 3 in comparison to ChatGPT 4.0 as an artificial intelligence tool in the diagnosis and therapy of primary head and neck cancer cases. European Archives of Oto-Rhino-Laryngology 2024. doi:10.1007/s00405-024-08828-1.
19. Loh E. ChatGPT and generative AI chatbots: challenges and opportunities for science, medicine and medical leaders. BMJ Leader 2024;8:51-54. doi:10.1136/leader-2023-000797.
20. Kataoka Y, Yamamoto-Kataoka S, So R, Furukawa TA. Beyond the Pass Mark: Accuracy of ChatGPT and Bing in the National Medical Licensure Examination in Japan. JMA J 2023;6:536-538. doi:10.31662/jmaj.2023-0043.
21. Huang RST, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung F-H. Assessment of Resident and AI Chatbot Performance on the University of Toronto Family Medicine Residency Progress Test: Comparative Study. JMIR Med Educ 2023;9:e50514. doi:10.2196/50514.
22. Plevris V, Papazafeiropoulos G, Jiménez Rios A. Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard; 2023. doi.org/10.3390/ai4040048