Section: Miscellaneous

Original Research Article

ROLE OF CHAT GPT 3.5 AND BARD AS SELF ASSESSMENT TOOL FOR UNDERGRADUATE LEVEL MULTIPLE CHOICE QUESTIONS IN OPHTHALMOLOGY

Abhijeet Khake[1], Sonali Khake[2], Pallavi Potdar[3], Ashutosh Potdar[4], Manjiri Desai[5]

[1]Assistant Professor, Department of Ophthalmology, SSPM Medical College, Sindhudurg, Maharashtra, India.
[2]Professor, Department of Anatomy, SSPM Medical College, Sindhudurg, Maharashtra, India.
[3]Professor, Department of Community Medicine, SSPM Medical College, Sindhudurg, Maharashtra, India.
[4]Professor, Department of Forensic Medicine, SSPM Medical College, Sindhudurg, Maharashtra, India.
[5]Assistant Professor, Department of Community Medicine, D Y Patil Medical College, Kolhapur, Maharashtra, India.

Received : *23/12/2023*
Received in revised form : *28/01/2024*
Accepted : *15/02/2024*

Keywords:
Artificial intelligence, deep learning, Chat GPT 3.5, Large Language Models (LLM), Multiple choice questions, Undergraduate medical education, Self assessment, Ophthalmology.

Corresponding Author:
Dr. Abhijeet Khake
Email: abhijeetkhake25@gmail.com

DOI:10.47009/jamp.2024.6.1.209

Source of Support: Nil,
Conflict of Interest: None declared

Int J Acad Med Pharm
2024; 6 (1); 1054-1058

Abstract

Background: Since 2015, significant progress has been made in the application of artificial intelligence (AI) particularly in ophthalmology. It has been used in identifying retinal problems based on fundus photographs and imaging. More recently interactive AI tools like Large Language Models (LLMs) are being explored in the domain of medical education and evaluation as it has shown to successfully pass medical licensing exam like USMLE. We have studied the self evaluation aspect of evaluation process in this study. **Aim:** To evaluate effectiveness of ChatGPT 3.5 & Google's Bard as a tool for self assessment of Multiple Choice Questions (MCQs) in ophthalmology for undergraduate students. **Material and Methods:** MCQs were selected from previous years question papers and available Competency based question and answer book for undergraduate in ophthalmology. Total of 137 questions were selected. Questions were segregated according to competencies as given in Competency based medical education (CBME) curriculum. Image based MCQs were excluded from the study. Single correct answer for each MCQ was identified and model answer key was prepared. Each question was asked to ChatGPT 3.5 and Google's Bard and the response were documented for both the AI tools. Correct response was scored as 1 and incorrect response was scored as 0. The responses were added for each topic and results tabulated for analysis. **Results:** Total of 137 MCQs were studied in this study which covered all the topics from ophthalmology as required for undergraduate students in CBME curriculum. Open AI's ChatGPT 3.5 gave 85 correct answers out of 137 i.e 62.04%. Google's Bard gave 72 correct responses out of 137 i.e 52.55%.**Conclusion:** We conclude that though both, Chat GPT 3.5 & Bard are above average in answering correct responses, the percentage of correct responses it can provide is not adequate in clinical branches like ophthalmology. Accuracy of >90% in MCQ's is expected to consider a particular tool reliable. Comparatively ChatGPT 3.5 has a better accuracy rate compared to Bard percentage wise. But this was not statistically significant. As both these tools provide comprehensive information about the multiple choice options it assists in making analytical choice amongst the options. In their current form, these tools are of limited use in self assessment by students. At best they can be used as tools to get quick information at preliminary stages.

INTRODUCTION

Since 2015, significant progress has been made in the application of artificial intelligence (AI) and deep learning (DL) in medicine, particularly in ophthalmology.[1] Deep learning has been widely used for image recognition using various types of ophthalmic data, such as fundus photographs and

OCT, and has shown strong results in detecting a wide range of diseases.[2,3] More recently, there has been growing interest in Natural Language Processing (NLP) in ophthalmology, which involves using AI to understand and interact with human language.[4] This has lead to development of Large Language Models (LLMs) like Open AI's ChatGPT and Google's Bard whose applications are been explored in ophthalmology. ChatGPT has been known to achieve near passing scores in medical licensing exam like USMLE.[5] Taking this into consideration, some researchers have studied the application of LLMs in medical education and explored their use in evaluation process of undergraduate students with mixed opinions. But studies related to evaluation for undergraduate students in ophthalmology are missing. We have addressed self-assessment aspect of undergraduate MCQs in ophthalmology in this study.

Present day undergraduate medical student, gains theoretical knowledge from books, didactic lectures (in person or online), small group teachings and briefings in clinical postings. Their knowledge and skills are assessed during regular exams. Theoretical assessment is done using long and short essays (which can be structured clinical questions or clinical reasoning questions) and MCQs. Whereas for self-assessment the students generally rely on books, previous years question papers and online resources.

With the advent of LLMs like ChatGPT and Google's Bard, it has made accessibility of relevant information more easy. While search engines like Google chrome can provide us with relevant websites to get information, LLMs can provide with relevant information directly, saving time and efforts. Moreover, LLMs are interactive and can also be used as an assessment tool. But as the currently available LLMs are not specifically designed for medical education in ophthalmology, we need to know their effectiveness as an assessment tool for the subject, before it can be used for self-assessment by undergraduate medical students. In this study we evaluate effectiveness of using Chat GPT and Google Bard as a self-assessment tool for MCQs in ophthalmology.

Aim

To evaluate effectiveness of Open AI's ChatGPT 3.5 & Google's Bard as a tool for self-assessment of MCQs in ophthalmology for undergraduate students.

MATERIALS AND METHODS

MCQs were randomly selected from previous years question papers and available Competency based question and answer book for undergraduate in ophthalmology.[6] Total of 137 questions were selected. Questions were segregated according to competencies as given in Competency Based Medical Education (CBME) curriculum by National Medical Commission of India (NMC). Image based MCQs were excluded from the study. Single correct answer for each MCQ was identified and model answer key was prepared. Recent edition of standard textbook of ophthalmology and Eyewiki an online ophthalmology website by American Academy of Ophthalmology were referred for formulating answers.[7] Each question was asked to ChatGPT 3.5 and Google's Bard and the response were documented for both the AI tools. Correct response was scored as 1 and incorrect response was scored as 0. The responses were added for each topic and results tabulated for analysis. Statistical analysis was done to know if one tool scores better than other. Mc Nemar test was used to determine the association of correct responses between the two LLMs. Statistical significance of $P \leq 0.05$ was considered significant.

RESULTS

Total of 137 MCQs were studied in this study which covered all the topics from ophthalmology as required for undergraduate students in CBME curriculum.

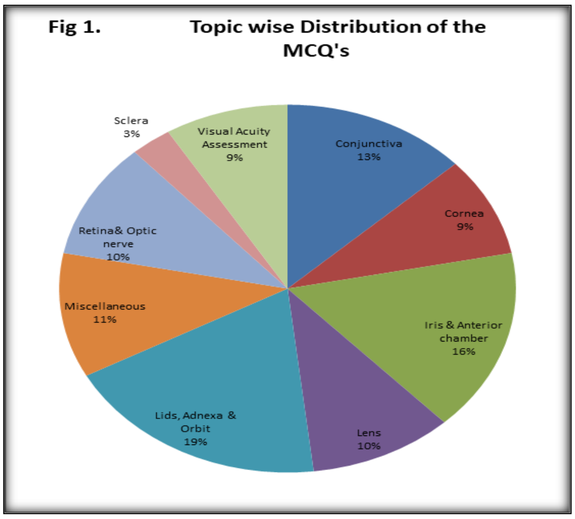

Figure 1: Topic wise Distribution of theMCQ's

Topic wise distribution is given in Fig 1. MCQs were categorised into 9 topics. Topic wise distribution of scores is given in Table 1. Three topics had same number of correct responses whereas for the topic of retina and optic nerve there was gross difference in the number of correct responses between ChatGPT and Bard, with ChatGPT being significantly superior.

Open AI's ChatGPT 3.5 gave 85 correct answers out of 137 i.e 62.04%. Google's Bard gave 72 correct responses out of 137 i.e 52.55%. But there is no statistically significant difference in total score between the two.

Table 1: Topic wise distribution of scores

Topic	No. of Questions	ChatGPT Score	Bard Score	P Value
		No.s of Correct answer/Total answers	**No.s of Correct answer/Total answers**	
Conjunctiva	18	11/18.	11/18.	0.72
Cornea	12	5/12.	5/12.	0.61
Iris & Anterior chamber	22	15/22.	12/22.	0.44
Lens	14	9/14.	8/14.	1.0
Lids, Adnexa & Orbit	26	13/26.	13/26.	1.0
Miscellaneous	15	10/15.	11/15.	1.0
Retina& Optic nerve	14	10/14.	5/14.	0.13
Sclera	4	3/4.	2/4.	1.0
Visual Acuity Assessment	12	9/12.	5/12.	0.22
TOTAL	137	85/137	72/137	0.11
% wise		62.04%	52.55%	

ChatGPT and Bard were compared with Mc Nemar test for different set of questions. The difference between responses given by ChatGPT and Bard is not statistically significant, P value is > 0.05.

Table 2: Topic wise 2 x 2 tables

	Conjunctiva		
	ChatGPT		
Bard	Yes	No	Total
Yes	7	4	11
No	4	3	7
Total	11	7	18

P value = 0.72 (not significant)

	Cornea		
	ChatGPT		
Bard	Yes	No	Total
Yes	3	2	5
No	2	5	7
Total	5	7	12

P value = 0.61 (not significant)

Iris & Anterior Chamber			
	ChatGPT		
Bard	Yes	No	Total
Yes	10	2	12
No	5	5	10
Total	15	7	22

P value = 0.44 (not significant)

	Lens		
	ChatGPT		
Bard	Yes	No	Total
Yes	5	3	8
No	4	2	6
Total	9	5	14

P value = 1.0 (not significant)

Lids, Adnexa & Orbit			
	ChatGPT		
Bard	Yes	No	Total
Yes	11	3	14
No	2	10	12
Total	13	13	26

P value = 1.0 (not significant)

	Miscellaneous		
	ChatGPT		
Bard	Yes	No	Total
Yes	8	3	11
No	2	2	4
Total	10	5	15

P value = 1.0 (not significant)

Retina & Optic Nerve			
	ChatGPT		
Bard	Yes	No	Total
Yes	4	1	5
No	6	3	9
Total	10	4	14

P value = 0.13 (not significant)

	Sclera		
	ChatGPT		
Bard	Yes	No	Total
Yes	2	0	2
No	1	1	2
Total	3	1	4

P value = 1.0 (not significant)

Visual Acuity Assessment			
	ChatGPT		
Bard	Yes	No	Total
Yes	4	1	5
No	5	2	7
Total	9	3	12

P value = 0.22 (not significant)

	Total		
	ChatGPT		
Bard	Yes	No	Total
Yes	54	19	73
No	31	33	64
Total	85	52	137

P value = 0.11 (not significant)

DISCUSSION

In our study we tried to find out if LLM's like Open AI's ChatGPT and Google's Bard can be used as a self-assessment tool for undergraduate level MCQ's in ophthalmology. We randomly selected 137 MCQs and asked them to Open AI's ChatGPT and Google's Bard. The responses were compared with model answer key. We found that ChatGPT gives correct answer 62% of time while Bard gives correct answer only 52% of times. But this difference was not statistically significant. Also a score of 62% is not an acceptable score particularly for clinical branches. More than 90% score is what we should be looking for which will make these tools reliable for self-study by undergraduate students. According to our study, with overall best accuracy rate being only 62%, which is for Chat GPT, it's better to limit its use to preliminary stages for quick & rough self-assessment by undergraduate medical students for Ophthalmology subject.

Agarwal M et al. in their study found that LLMs like Chat GPT, Bard and Bing can generate assessment questions of varying difficulty in the subject of Physiology but are not yet fully developed and had its own limitations.[8]

Das D et al. studied efficacy of ChatGPT's responses to 1st order and 2nd order knowledge question based on CBME guidelines for subject of microbiology. They studied total of 96 questions. They found that ChatGPT achieved accuracy of around 80% with no difference in answering 1st order and 2nd order knowledge questions and concluded that ChatGPT is a effective tool to answer these questions in microbiology.[9] Sinha RK et al. studied applicability of ChatGPT in solving higher order reasoning questions in the subject of Pathology. They found that ChatGPT scored around 80% accuracy while answering 100 high order questions.[10] Ghosh et al. studied ChatGPTs ability to solve higher order questions in subject of Biochemistry. They studied 200 questions out of which 100 were reasoning type questions which required higher order thinking and found that ChatGPT scored atleast 4 out of 5 marks in all questions i.e 75%.[11]

Surapaneni KM et al. also evaluated ChatGPT as a self-learning tool in medical biochemistry by asking Chat GPT to solve questions from exam paper. They studied 100 MCQs and found that Chat GPT provided relevant and appropriate answers to all the MCQs i.e 100%. They concluded that ChatGPTs overall score was only 58% and needs improvement.[12] However in our study the percentage of correct responses for MCQs provided by Chat GPT was only 62%. This may be due to a greater need of analytical and logical thinking required to answer questions in ophthalmology.

We studied a total of 137 MCQs which gave us best accuracy rate of 62% with Chat GPT. If a larger sample size is studied or if topic wise study with larger sample size is studied, it can give a more accurate and categorized results which can give a better understanding. But even with the results from this study we can conclude that use of LLMs should be limited to preliminary stages for quick & rough self-assessment by undergraduate medical students for Ophthalmology subject. We hope in future these tools will refined further and provide more that 90% accuracy rate or a specialized LLM will be developed for medical field which can give near accurate responses.

Role of AI is going to increase in medical field in future and we need to consider this and align medical education accordingly. There has always been a big question regarding when AI should be introduced in medical curriculum. As undergraduate curriculum is jam-packed, with the main focus on gaining knowledge and skill to become a compassionate doctors with critical thinking, the general consensus to introduce full-fledged AI is during post graduate/fellowship training, when students basic clinical knowledge, skills, and understanding of clinical decision-making and workflow are more developed.[13] However as LLMs are a simpler form of AI tools which do not need additional training and efforts to use them, these can be considered as one of the earliest tools to introduce AI in medical education in the form of self-assessment tool in early stages of topic learning for undergraduate students.

CONCLUSION

In this study we conclude that though both, Chat GPT 3.5 & Bard are above average in answering correct responses, the percentage of correct responses it can provide is not adequate in clinical branches like ophthalmology. Comparatively percentage-wise ChatGPT 3.5 has a better accuracy rate compared to Bard. As both these tools provide comprehensive information about the multiple choice options, it assists in making a more informed and analytical choice amongst the options. In their current form, these tools are of limited use in self assessment by students. At best they can be used as tools to get quick information at preliminary stages.

REFERENCES

1. Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol.2019; Feb;103(2):167-175.
2. Ursula Schmidt-Erfurth, Amir Sadeghipour, Bianca S Gerendas, Sebastian M Waldstein, Hrvoje Bogunović. Artificial intelligence in retina. Prog Retin Eye Res. 2018 Nov: 67:1-29
3. Fares Antaki, Razek Georges Coussa, Ghofril Kahwati, Karim Hammamji, Mikael Sebag, Renaud Duval. Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images. Br J Ophthalmol. 2023; Jan;107(1):90-95.
4. Siddharth Nath, Abdullah Marie, Simon Ellershaw, Edward Korot, Pearse A Keane. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol. 2022 Jul;106(7):889-892.
5. Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2): e0000198
6. Madhurima Nayak, Sushrutha Academy. Competency based question and answer in Ophthalmology for 3rd MBBS Professional exam. CBS Publishers & Distributors Pvt. Ltd.
7. Ramanjit Sihota, Radhika Tandon, Parsons Diseases of the eye 24th edition, Elsevier Publications
8. Agarwal M, Sharma P, Goswami A. Analysing the Applicability of ChatGPT, Bard, and Bing to Generate Reasoning-Based Multiple-Choice Questions in Medical Physiology. Cureus. 2023 Jun 6;15(6): e40977.
9. Dipmala Das, Nikhil Kumar, Langamba Angom Longjam, Ranwir Sinha, Asitava Deb Roy, Himel Mondal, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus. 2023 Mar 12;15(3): e36034
10. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. Cureus 2023 Feb 20;15(2): e35237
11. Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023 Apr 2;15(4): e37023.
12. Surapaneni KM, Rajajagadeesan A, Goudhaman L, Lakshmanan S, Sundaramoorthi S, Ravi D, et al. Evaluating ChatGPT as a self-learning tool in medical biochemistry: A performance assessment in undergraduate medical university examination. Biochemistry and Molecular Biology Education. 19 Dec 2023. https//doi.org/10.1002/bmb.21808
13. Ngo B, Nguyen D, vanSonnenberg E. The Cases for and against Artificial Intelligence in the Medical School Curriculum. Radiol Artif Intell. 2022 Aug 17;4(5): e220074.