Section: Miscellaneous

Original Research Article

ROLE OF CHATGPT 3.5 AND BARD AS SELF ASSESSMENT TOOL FOR UNDERGRADUATE LEVEL LONG ANSWER QUESTIONS IN OPHTHALMOLOGY

Abhijeet Khake[1], Suvarna Gokhale[2], Pradeep Dindore[3], Sonali Khake[4], Pallavi Potdar[5], Ashutosh Potdar[6], Manjiri Desai[7]

[1]Assistant Professor, Department of Ophthalmology, SSPM Medical College, Sindhudurg, Maharashtra, India
[2]Ex HOD & Professor, Department of Ophthalmology, Smt. Kashibai Navale Medical College & General Hospital, Pune, Maharashtra, India
[3]Professor, Department of Ophthalmology, Prakash Institute of Medical Sciences & Research, Islampur, Maharashtra, India
[4]Professor, Department of Anatomy, SSPM Medical College, Sindhudurg, Maharashtra, India
[5]Professor, Department of Community Medicine, SSPM Medical College, Sindhudurg, Maharashtra, India
[6]Professor, Department of Forensic Medicine, SSPM Medical College, Sindhudurg, Maharashtra, India
[7]Assistant Professor, Department of Community Medicine, D Y Patil Medical College, Kolhapur, Maharashtra, India.

Received : 03/01/2024
Received in revised form : 24/02/2024
Accepted : 12/03/2024

Keywords:
Artificial intelligence, deep learning, ChatGPT 3.5, Large Language Models (LLM) , Long Answer Questions (LAQ), Undergraduate education, Self-assessment, Ophthalmology.

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

DOI: 10.47009/jamp.2024.6.2.56

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

Int J Acad Med Pharm
2024; 6 (2); 264-269

Abstract

Background: Use of artificial intelligence (AI) has been gradually increasing in medical field. It has also found multiple applications in the field of ophthalmology. It has been used in diagnosis and grading of common retinal pathologies like diabetic retinopathy and glaucoma. In last few years, Large Language Models (LLMs) which is an interactive AI tool, are being investigated for its applicability in medical education. LLMs have also been studied as an evaluation tool for the subjects of physiology, biochemistry, pathology and microbiology. In this study we tried to find out how effective are these tools in self-assessment of undergraduate level long answer questions in ophthalmology. **Aim:** To evaluate effectiveness of ChatGPT 3.5 & Google Bard as a tool for self-assessment of Long Answer Questions (LAQs) in ophthalmology for undergraduate students. **Material and Methods:** LAQs were selected from previous question papers and available question banks in ophthalmology. Total of 35 LAQs with 4 questions each were randomly selected from the question pool. All long answer questions were segregated according to competencies as given in Competency Based Medical Education (CBME) curriculum of National Medical Commission of India (NMC). Image based LAQs were excluded from the study. Model answers were prepared for all these questions by 3 ophthalmologists with mutual discussion. Each question was then asked to ChatGPT 3.5 and Google Bard and the responses were evaluated and graded on a scale of 3, for correct diagnosis, content accuracy and relevance for both the AI tools by 3 ophthalmologists. The responses were evaluated for each topic and results tabulated for analysis. **Results:** Total of 35 LAQs were studied in this study which covered all the topics from ophthalmology as required for undergraduate students in CBME curriculum. The total score for accurateness & adequateness of information provided by ChatGPT and Bard is 68 out of 105 i.e 64.76%. Important aspects of a topic like staging, classification, specific treatment, orderly steps of surgery are not specific or missed many a times by both. Both missed diagnosis in 4 out of 35 questions i.e 11% of times. **Conclusion:** In this study though ChatGPT and Bard can provide answers to LAQs their answers cannot be relied on with confidence every time. Using ChatGPT or Bard alone for self-assessment while studying for LAQs by undergraduate students is not advisable. Students should use standard books or standard online resources for ophthalmology while preparing for LAQ problems.

INTRODUCTION

Artificial intelligence (AI) and deep learning (DL) are being explored in medicine since 2015, especially in the subject of Ophthalmology.[1] Deep learning has shown promising results for detecting retinal pathologies from fundus photographs and OCT images especially in diabetic retinopathy and glaucoma. [2,3] This has translated into newer AI based diagnostic tools. Natural Language Processing (NLP), a branch of AI dealing with understanding and interacting with human language has gained particular interest in ophthalmology.[4] This has led to development of Large Language Models (LLMs) like Open AI's ChatGPT and Google's Bard. Application of these LLMs is being investigated in medical education. Recently there were reports of ChatGPT having successfully passed medical licensing exam like USMLE.[5] Encouraged by these developments, researchers have explored use of LLMs to evaluate undergraduate medical students with mixed opinions. LLMs have been studied as a evaluation tool for physiology, biochemistry, microbiology and pathology for undergraduate students. But their similar use in ophthalmology is not studied yet. In this study we have addressed the self-assessment aspect of evaluation process for undergraduate level long answer questions (LAQ) using these tools.

Now a days students acquire knowledge from multiple sources which include 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 answer questions (LAQ) and short answer questions (SAQ), which can be structured clinical questions or clinical reasoning questions, and MCQs. LAQs are clinical scenario based, image based, audio based or video based problems with generally 4 to 5 questions related to the problem which are used to assess knowledge and understanding of different aspects of a topic. As per Competency Based Medical Education (CBME) guidelines, LAQs should pose a clinical/practical problem to the students and require them to apply knowledge and integrate it with disciplines. The crucial aspect of LAQs is to get the diagnosis of the given problem correctly. Without correct diagnosis the questions posed in the problem may not be answered correctly. For LAQs, students use books, previous years question papers and online resources for self-study and self-assessment

With the introduction of LLMs like ChatGPT and Google Bard, accessing relevant information has become very 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. The other advantage of LLMs is that they are interactive and hence can be used as an assessment tool. But as LLMs are still in developing stage, LLMs specifically for medical education in ophthalmology are not yet developed. So we need to know if currently available LLMs are good enough to be used as assessment tool, before it can be used for self study and self-assessment by undergraduate medical students. In this study we evaluate effectiveness of using ChatGPT and Google Bard as a self-assessment tool for LAQs in ophthalmology.

Aim

To evaluate effectiveness of ChatGPT 3.5 & Google Bard as a tool for self-assessment of Long Answer Questions (LAQs) in Ophthalmology for undergraduate students.

MATERIAL AND METHODS

Long answer questions (LAQ) were selected from previous question papers and available question banks in ophthalmology. For the purpose of this study only Clinical based scenarios were included in the study. Total of 35 structured LAQs were randomly selected from the question pool. Questions were segregated according to competencies as given in CBME curriculum of National Medical Council of India (NMC). Image based Long answer questions were excluded from the study. Each LAQ had 4 sub questions. Model answers were prepared for all these questions by 3 subject experts in ophthalmology with mutual discussion. Standard textbook of ophthalmology and Eyewiki an online ophthalmology website by American Academy of Ophthalmology were referred for preparing model answers.[6] Each question was then asked to ChatGPT 3.5 and Google Bard and the responses were evaluated for correct diagnosis, content accuracy and relevance for both the AI tools by 3 ophthalmologists. Correctness of diagnosis for each question was noted for both AI tools. No response for any question by an AI tool was considered as wrong diagnosis. For assessing content accuracy, adequateness and relevance of answer provided for each question, the responses were scored on the scale of 3 according to scoring system given below. 3 responses from 3 ophthalmologists for each question were noted for each AI tool. These 3 responses were then converted into 1 single response. This was achieved by selecting the score given by minimum 2 ophthalmologists which was same. This was taken as final score for that particular question for that AI tool. 3 different score from 3 ophthalmologists was not observed as model answers which were prepared before starting evaluation of the tools were prepared with mutual discussion by the 3 subject experts in ophthalmology.

Scoring of the responses was done as below

1. **Not matching with standard answer. (Bad Answers)**

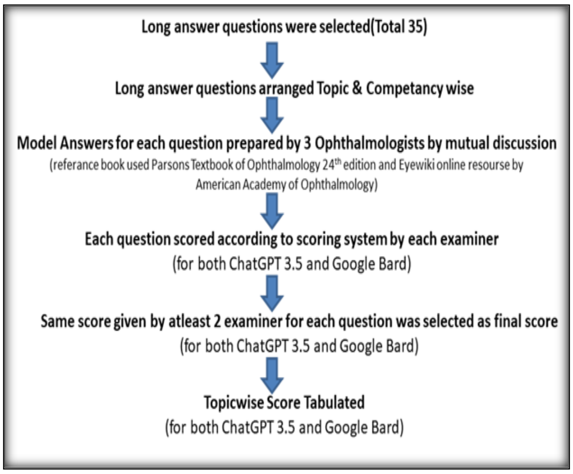
This will include answers where diagnosis is wrong or no answer, concept is wrong, gross deviation from established concepts and knowledge AND/OR <50% key point covered

2. **Partial match with standard answer. (Average acceptable answers)**

This will include answers with correct concept, but critical/important points not covered AND/OR key points covered between the range of 50% to 80%

3. **Complete match with standard answer. (good answers)**

This will include answers with correct concept, with critical/important points covered AND/OR more than 80% of key points covered.



Long answer questions were selected(Total 35)

Long answer questions arranged Topic & Competancy wise

Model Answers for each question prepared by 3 Ophthalmologists by mutual discussion
(referance book used Parsons Textbook of Ophthalmology 24th edition and Eyewiki online resourse by American Academy of Ophthalmology)

Each question scored according to scoring system by each examiner
(for both ChatGPT 3.5 and Google Bard)

Same score given by atleast 2 examiner for each question was selected as final score
(for both ChatGPT 3.5 and Google Bard)

Topicwise Score Tabulated
(for both ChatGPT 3.5 and Google Bard)

Figure 1: Flow chart of process followed

All these values were then tabulated in an excel sheet and evaluated. Statistical analysis was done using the SPSS software. Z test for proportion was applied to know if one tool scores better than other.

RESULTS

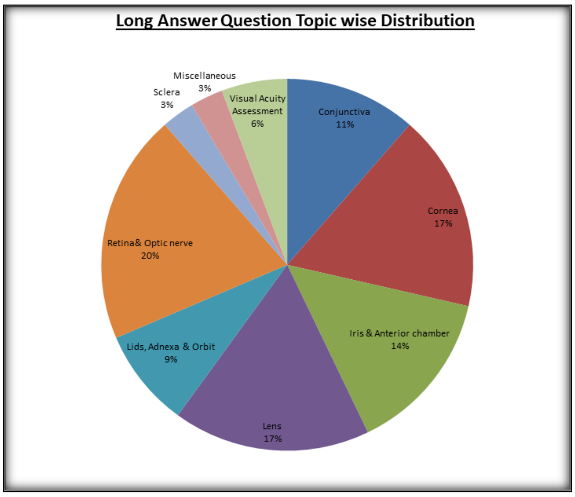


Figure 2: Long Answer Question Topic wise Distribution

Total of 35 LAQs were studied and graded on a scale of 3. Topic wise percentage distribution is shown in the above Fig 2. The maximum total score is 105. The detailed topic wise analysis is shown in Table 1 to 6 below

As LAQs are based on the clinical case scenarios, if the diagnosis goes wrong the further answers to the questions asked in the problem go wrong. Both ChatGPT & Bard provided correct diagnosis for 31 questions out of 35 questions i.e around 88.57%. ChatGPT provided wrong diagnosis for 4 question whereas Bard gave wrong diagnosis to 2 questions and no diagnosis to 2 questions. These diagnosis were included under bad answer. For statistical analysis SPSS software was used and Z test for proportion was applied. Statistical analysis showed no statistically significant different between ChatGPT and Bard in getting the diagnosis correct, overall as well as topic wise. P value $\leq$ 0.05 was considered as statistical significant. Details are shown in Table 1 & 2.

Table 3 shows the detailed content analysis and Table 4 shows content score analysis. The total score for accurateness & adequateness of information provided by ChatGPT and Bard is 68 out of 105 i.e 64.76%. Statistical analysis showed no statistically significant different between ChatGPT and Bard in providing accurate & adequate content, overall as well as topic wise. P value $\leq$ 0.05 was considered as statistical significant. Both ChatGPT and Bard produce responses but with average accuracy. Also it was observed that, important aspects of a topic like staging, classification, specific treatment, orderly steps of surgery are not specific or missed by both in many answers.

Responses with wrong diagnosis, inadequate or grossly wrong content were considered as bad responses not suitable for undergraduate student learning. Bad responses in case of ChatGPT is 5 out 35 which is approximately 14% whereas Bard provided wrong diagnosis for 2 questions, no diagnosis for 2 questions and grossly wrong/inadequate content answers to 9 questions, a total of 13 out of 35 bad responses i.e around 37%.

Table 5 & Fig 3 shows over all content quality analysis and Table 6 & Fig 4 shows topic wise content quality analysis. When questions with bad answers are excluded i.e only good answers and average acceptable answers are considered , ChatGPT provides responses to 30 questions out of 35 i.e around 86% which are content wise accurate and adequate. Bard provides responses with relevant and adequate content in 22 out of 35 questions i.e 63%. Out of 35 questions ChatGPT provides good answers to only 7 questions i.e approximately 20% of times as compared to Bard which provides good answers to 11 questions i.e approximately 31% of time. Though Bard has better chance than ChatGPT in providing good answers, the overall content accuracy and adequateness provided by ChatGPT is better than Bard when questions are selected. But this sort of selection of questions cannot be done beforehand in actual practice as the pool of questions is unlimited.

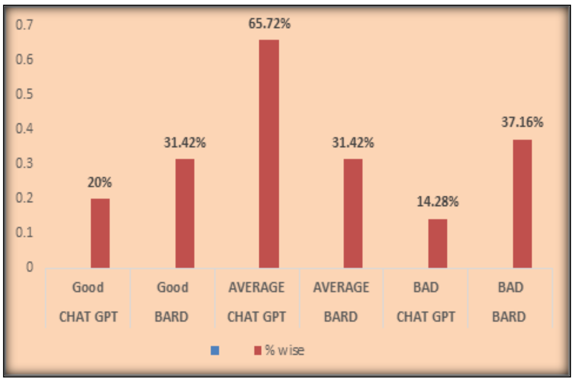
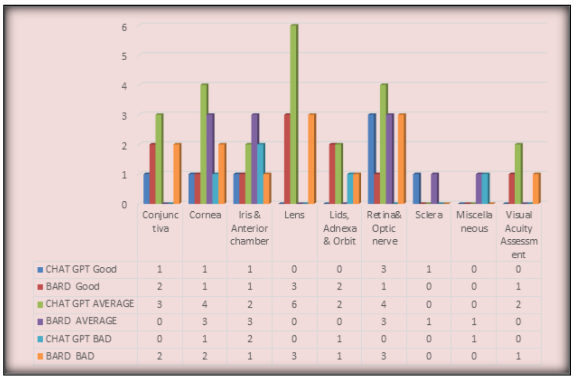


Figure 3: Content Quality Analysis



	Conjunc tiva	Cornea	Iris & Anterior chamber	Lens	Lids, Adnexa & Orbit	Ret ina& Optic nerve	Sclera	Miscella neous	Visual Acuity Assessm ent
■ CHAT GPT Good	1	1	1	0	0	3	1	0	0
■ BARD Good	2	1	1	3	2	1	0	0	1
■ CHAT GPT AVERAGE	3	4	2	6	2	4	0	0	2
■ BARD AVERAGE	0	3	3	0	0	3	1	1	0
■ CHAT GPT BAD	0	1	2	0	1	0	0	1	0
■ BARD BAD	2	2	1	3	1	3	0	0	1

Figure 4: Content Quality Analysis - Topic wise

Table 1: Diagnosis analysis

Topic	No. of Questions	Diagnosis Analysis			
		Correct Diagnosis		Wrong Diagnosis	
		ChatGPT	Bard	ChatGPT	Bard
Conjunctiva	4	4	4	0	0
Cornea	6	5	5	1	1
Iris & Anterior chamber	5	3	5	2	0
Lens	6	6	6	0	0
Lids, Adnexa & Orbit	3	2	3	1	0
Retina& Optic nerve	7	7	4	0	1 (2 no ans)
Sclera	1	1	1	0	0
Miscellaneous	1	1	1	0	0
Visual Acuity Assessment	2	2	2	0	0
TOTAL	35	31	31	4	4
% wise		88.57%	88.57%	11.43%	11.43%

Table 2: Correct Diagnosis Analysis

Topic	No. of Questions	Correct Diagnosis ChatGPT	Correct Diagnosis Bard	p value
Conjuctiva	4	4	4	1
Cornea	6	5	5	1
Iris & Anterior chamber	5	3	5	0.11
Lens	6	6	6	1
Lids, Adnexa & Orbit	3	2	3	0.27
Retina& Optic nerve	7	7	4	0.51
Sclera	1	1	1	1
Miscellaneous	1	1	1	1
Visual Acuity Assessment	2	2	2	1
TOTAL	35	31	31	1
% wise		88.57%	88.57%	

There is no statistically significant difference between ChatGPT and Bard for correct diagnosis analysis for different topics (P value > 0.05).

Table 3: Content Analysis

Topic	No. of Questions	Content Analysis							
		Chat GPT Score	Bard Score	ChatGPT Answers			Bard Answers		
		Total score/Max score	Total score/Max score	Goo d	Avg	Bad	Good	Avg	Bad
Conjunctiva	4	9/12.	8/12.	1	3	0	2	0	2
Cornea	6	12/18.	11/18.	1	4	1	1	3	2
Iris & Anterior chamber	5	9/15.	10/15.	1	2	2	1	3	1
Lens	6	12/18.	12/18.	0	6	0	3	0	3
Lids, Adnexa & Orbit	3	5/9.	7/9.	0	2	1	2	0	1
Retina& Optic nerve	7	13/21	12/21.	3	4	0	1	3	3
Sclera	1	3/3.	2/3.	1	0	0	0	1	0
Miscellaneous	1	1/3.	2/3.	0	0	1	0	1	0
Visual Acuity Assessment	2	4/6.	4/6.	0	2	0	1	0	1
TOTAL	35	68/105	68/105	7	23	5	11	11	13
% wise		64.76%	64.76%	20 %	65.72 %	14.28 %	31.42 %	31.42 %	37.16 %

Table 4: Content Score Analysis

Topic	Content Analysis Chat GPT Score	Content Analysis Bard Score	p value
	Total score/Max score	Total score/Max score	
Conjunctiva	9/12.	8/12.	0.65
Cornea	12/18.	11/18.	0.72
Iris & Anterior chamber	9/15.	10/15.	0.7
Lens	12/18.	12/18.	1
Lids, Adnexa & Orbit	5/9.	7/9.	0.31
Retina& Optic nerve	13/21	12/21.	0.75
Sclera	3/3.	2/3.	0.27
Miscellaneous	1/3.	2/3.	0.41
Visual Acuity Assessment	4/6.	4/6.	1
TOTAL	68/105	68/105	1
% wise	64.76%	64.76%	

There is no statistically significant difference between ChatGPT and Bard for content score analysis for different topics (P value > 0.05).

Table 5: Content Quality Analysis

	ChatGPT	Bard	ChatGPT	Bard	ChatGPT	Bard
Content Quality	**Good**	**Good**	**Average**	**Average**	**Bad**	**Bad**
% wise	20%	31.42%	65.72%	31.42%	14.28%	37.16%

Table 6: Content Quality Analysis – Topic wise

Topic	ChatGPT	Bard	Chat GPT	Bard	Chat GPT	Bard
	Good	**Good**	**Average**	**Average**	**Bad**	**Bad**
Conjunctiva	1	2	3	0	0	2
Cornea	1	1	4	3	1	2
Iris & Anterior chamber	1	1	2	3	2	1
Lens	0	3	6	0	0	3
Lids, Adnexa & Orbit	0	2	2	0	1	1
Retina& Optic nerve	3	1	4	3	0	3
Sclera	1	0	0	1	0	0
Miscellaneous	0	0	0	1	1	0
Visual Acuity Assessment	0	1	2	0	0	1
TOTAL	7	11	23	11	5	13
% wise	20%	31.42%	65.72%	31.42%	14.28%	37.16%

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 LAQ's in ophthalmology. We randomly selected 35 structured LAQs and asked them to Open AI's ChatGPT and Google's Bard. Each LAQ had 4 sub questions. The responses were compared for relevance and accuracy of information with model answer key, and scored on a scale of 3, by 3 Ophthalmologists. The responses were graded into good responses, average acceptable responses and bad responses. We found that total score for accurateness & adequateness of information provided by ChatGPT and Bard it is around 65% suggesting that both ChatGPT and Bard can provide responses with average accuracy. Important aspects of a topic like staging, classification, specific treatment, orderly steps of surgery are not specific or missed by both in many answers. Both ChatGPT & Bard gave wrong diagnosis 11% of times. ChatGPT provided good answers only 20% of times as compared to Bard which was around 31% of time. Grossly wrong answers which are not suitable for undergraduate students learning, provided by ChatGPT is around 14% as compared to Bard which is 37%. Comparatively Bard has better chance than ChatGPT in providing good answers, but the overall content accuracy and adequateness provided by ChatGPT is better than Bard if questions are selected. During self assessment as the student does not have option of selecting question and the pool of questions is unlimited. Both these tools cannot be relied on with confidence every time for correctness and adequateness of responses provided by these AI tools to these questions. Using ChatGPT or Bard alone for self-assessment while studying for LAQs by undergraduate students is not advisable.

Agarwal M et al studied if LLMs like ChatGPT, Bard and Bing can generate assessment questions in Physiology. They found that LLMs can generate assessment questions of varying difficulty levels but they have their own limitations. They concluded that these LLMs need to develop further for their effective use.[7]

Das D et al studied 1st order and 2nd order knowledge question using ChatGPT in the subject of microbiology. Total of 96 questions were studied. They observed no difference by ChatGPT in answering 1st order and 2nd order knowledge question. The accuracy achieved in their study was around 80%. They concluded that ChatGPT is an effective tool to answer these questions in microbiology.[8] Sinha R K et al. studied higher order reasoning questions using ChatGPT in the subject of

Pathology. They found that ChatGPT answered high order questions with around 80% accuracy.[9] Ghosh et al. studied higher order questions from the subject of Biochemistry using ChatGPT. They studied 100 reasoning type questions which required higher order thinking and found that ChatGPT scored more than 75% score in all questions.[10] In our study we achieved around 86% accuracy with ChatGPT and 63% with Bard. But important aspects of a topic like staging, classification, specific treatment, orderly steps of surgery are not specific or missed by both in many answers. Also ChatGPT had 14% chance as compared to Bard which had 37% chance of providing grossly wrong answers which are not suitable for undergraduate student learning. This may be because of more analytical thinking required to answer questions in clinical subjects like ophthalmology.

Surapaneni K M et al. explored ChatGPT as a self-learning tool in medical biochemistry. They used ChatGPT to solve questions from exam paper. The overall score ChatGPT achieved was only 58%. They concluded that this score was not good enough and needs improvement for which ChatGPT must focus on generating not only accurate but also comprehensive and contextually relevant content. Only then ChatGPT can be used as a self-learning tool by undergraduate students.[11] In our study we also concluded that for LAQs these AI tools in their current form cannot be relied on with confidence for all questions posed. There is still lot of work to be done to bring them to reliable use.

CONCLUSION

We conclude that though ChatGPT and Bard can provide correct answers to LAQs, their answers cannot be relied on with confidence all the time and their use in answering LAQs for self-assessment by undergraduate students in ophthalmology is not advisable. Students should use standard text books or standard online resources for ophthalmology while preparing for LAQ problems.

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. Ramanjit Sihota, Radhika Tandon, Parsons Diseases of the eye 24th edition, Elsevier Publications
7. 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.
8. 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
9. 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
10. 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.
11. 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.