Assessing ChatGPT, Bing, Bard's competency in Neuro-ophthalmology : A Textbook Review Questions Trial

Mert Egemen Caliskan[1], Semih Ceylan[1] ,Musacan Yalcin[1], Betin Bilkan Karaman[1], Eray Atalay[2]

[1]Eskisehir Osmangazi University Medical School, [2]Department of Ophthalmology, Eskisehir Osmangazi University Faculty of Medicine

Introduction:

This study aims to assess the competency of three popular open access Large Language Models (LLM) Chatbots, namely ChatGPT 3.5 (Open AI), Bing (Microsoft) and Bard (Google), in answering 64 ''Review Questions'' from the Neuro-ophthalmology section of the Review of Ophthalmology textbook (Elsevier, 4th edition).

Methods:

- All questions were multiple choice with one choice being correct. Questions were asked individually on a new chat window each time.
- The prompt structure was *''The Question with Choices''* + *''What is the correct answer?''*.
- The answers were contrasted with the answer sheet of the textbook and labelled as correct or incorrect. Questions were categorized as "neuroanatomy", "physiology" and "clinical" as per the content in the book.
- Statistical analyses were performed using Chi-square test and statistical significance was set at p <0.05.



10th of December 2023	ChatGPT (OpenAI)			Bing (Microsoft Copilot)			Bard (Google Gemini)		
	Correct	Incorrect	Pct.	Correct	Incorrect	Pct.	Correct	Incorrect	Pct.
Neuroanatomy	5	1	83%	5	1	83%	4	2	66%
Physiology	4	4	50%	5	3	62%	5	3	62%
Clinical	26	14	65%	23	17	57%	24	16	60%

Models were tested on 10th of December 2023.

Friedman, N.J., P.K. Kaiser, and W.B. Trattler, Review of Ophthalmology. 2022: Elsevier.

Results:

- Among 64 questions, ChatGPT 3.5 performed best answering 35 questions correct (54.7%), followed by Bard and Bing both answering 33 questions correct (51.6%). However, the difference was not statistically significant (p= 0.91).
- For 6 neuroanatomy questions, ChatGPT 3.5 and Bing answered 5 questions correct, Bard answered 4 questions correct (p= 0.72).
- For physiology ChatGPT 3.5 performed the worst answering 4 questions correct out of 8, Bard and Bing performed better both answering 5 out of 8 questions correct (p= 0.84).
- For clinical, ChatGPT 3.5 performed the best answering 26 out of 50 questions correct followed by Bard answering 24 out of 50 correct and lastly Bing answering 23 out of 50 questions correct (p= 0.82).

Conclusions:

LLM Chatbots were only able to answer approximately half of the questions correctly. The chatbots performed similarly for different question categories.