“Comparative analysis of large language models (LLMs) in the NHS 111 online triaging for Emergency ophthalmology”

Shaheryar Ahmed Khan[1], Chrishan Gunasekera[2]
Colchester Hospital, East Suffolk and North Essex NHS Foundation Trust[1] , Norfolk & Norwich University Hospital[2]

BACKGROUND

In recent years, artificial intelligence (AI) has demonstrated remarkable capabilities in various applications, including healthcare [1, 2]. Ophthalmology has seen a particular increase in the use of AI [3]. In this study, we assess the performance of different LLM-chatbots in comparison to the ‘NHS online 111’ triage system for ophthalmology emergency responses

METHODS

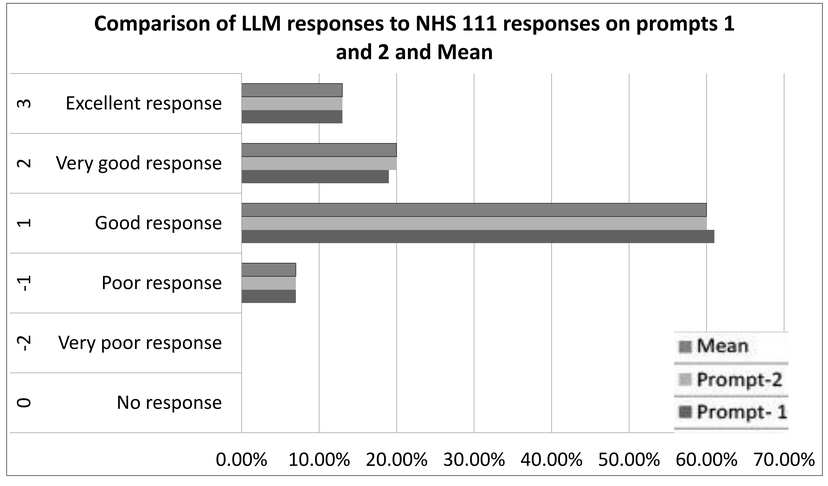
We included 21 ophthalmology related emergency scenario questions from the 111 triaging question algorithm. These questions were based on four different ophthalmology emergency themes . We first utilized the NHS 111 online for answering each of the 21 ophthalmology emergency question as an adult patient. The responses generated from NHS 111 online, thereafter was compared to the different LLM-chatbots responses. For this purpose, we selected and included a range of state-of-the-art Large language models including ChatGPT-3.5, Google Bard, Bing Chat, and ChatGPT-4.0. We asked LLM chatbots the same identical question by feeding the question into each LLM chatbot individually. To evaluate the accuracy of responses this was done twice using the two prompt strategy. First prompt stating ‘as a patient living in the UK’ at the end of the question and second prompt stating ‘Tell me as if you are NHS 111 triage service’

The following grading system was devised and used for assessing the accuracy:**–2 graded as “Very poor”; –1 graded as “Poor”; 0 graded as “No response”; 1 graded as “Good”; 2 graded as “Very good”:;3 graded as “Excellent”**

TABLE 1- Comparison of Mean score of overall ‘LLM’ responses to ‘NHS 111 online’ responses across all four themes of questions

Scoring	Grading	Theme 1 questions	Theme 2 questions	Theme 3 questions	Theme 4 questions
-2	Very poor response	0%	0%	0%	0%
-1	Poor response	2%	2%	10.5%	13%
0	No response	0%	0%	0%	0%
1	Good response	60.5%	60.5%	62.5%	50.5%
2	Very good response	25%	25%	18.5%	17.5%
3	Excellent response	12.5%	12.5%	8.5%	19%

Figure 1



RESULTS

•The minimum optimum score chosen in comparing the responses of ‘NHS 111 online’ and LLMs’ for assessing the accuracy of LLMs’ in this study was “1” (Good)

•Of all the 21 questions answered by all the LLMs’ included, an overall score of ≥1 graded as “Good” or better was achieved by 93% responses of all LLMs (FIGURE-1). This refers to at least part of this answer having correct information and partially matching the NHS 111 response, as well as absence of any wrong information, response or advice which is potentially dangerous or harmful to the patient’s eye or health

•The overall LLM responses which scored “2” or ‘Very good’ (meaning wholly correct answer and matching almost similar to the ‘NHS 111 response’) was 20%, while LLMs’ scoring “3” or ‘Excellent’ (meaning complete and fully correct answers matching the NHS 111 responses but additionally containing detailed additional information useful in the response for patients and not containing any misinformation or potentially hazardous advise) was 13%. Collectively the score of >2 or better was achieved by 33% of all LLMs’. 7% of all the LLM responses were scored as “-1” (Poor). There were no responses which were scored as 0 (No response) or -2(Very poor) in this study.

•There was no marked difference and rather very similar results were seen overall on both the prompts. Overall, the mean accuracy score of >1 or better was achieved by 93% of LLMs’ and this was the same on both the prompts.

•Analyzing the answers across the four question themes, >1 or better score was achieved by 98% of overall LLM responses for both Theme 1 and Theme 2 questions, 89.5% for Theme 3 questions and 87% for Theme 4 questions (Table 1)

DISCUSSION & CONCLUSION

•The results indicate a high level of accuracy of LLM responses, with 93% achieving a score of "Good" or better when compared to the NHS 111 online . Furthermore, the absence of responses graded as "-2" (Very poor), indicating potentially harmful advice in this study is reassuring and demonstrates the safety of utilizing LLMs in such scenarios

•In conclusion, this study provides valuable insights into the performance of LLM-chatbots in emergency ophthalmology triage. The high accuracy and safety observed in LLM responses support their potential as effective tools for providing timely information and guidance to patients.

•While further research is warranted to validate these findings in clinical practice, LLMs hold promise in enhancing patient care and healthcare accessibility in the digital age

REFERENCES

1. Bates DW, Levine D, Syrowatka A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med*. 2021;4(1):54. doi: 10.1038/s41746-021-00423-6

2. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: A report from the national academy of medicine. *JAMA*. 2020;323(6):509–510. doi: 10.1001/jama.2019.21579

3. Jiang X, Xie M, Ma L, Dong L, Li D. International publication trends in the application of artificial intelligence in ophthalmology research: An updated bibliometric analysis. Ann Transl Med. 2023;11:219