Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

Assessing the Performance of Large Language Models in Maxillofacial Trauma Triage: A Prospective Observational Study

[1] Dr. Sejal Doshi [*], [2] Dr. Dyna Albert, [3] Dr. Murugesan K., [4] Dr. Vinod Krishna, [5] Dr. Santhosh Kumar P.

[1] Postgraduate Student, Department of Oral and Maxillofacial Surgery, Saveetha Dental College, Chennai, India
[2] Associate Professor, Department of Oral and Maxillofacial Surgery, Saveetha Dental College, Chennai, India
[3] Professor and Head, Department of Oral and Maxillofacial Surgery, Saveetha Dental College, Chennai, India
[4] Assistant Professor, Department of Oral and Maxillofacial Surgery, Saveetha Dental College, Chennai, India
[5] Professor, Department of Oral and Maxillofacial Surgery, Saveetha Dental College, Chennai, India

(Received: 16 July 2025 Revised: 20 August 2025 Accepted: 20 September 2025)

KEYWORDS

Maxillofacial trauma, triage, ChatGPT4o, artificial intelligence.

ABSTRACT:

Background: The integration of artificial intelligence (AI) and related technologies into healthcare is on the rise. In trauma triage, effective decision-making plays a crucial role in managing mass casualty incidents. This study aims to evaluate the clinical accuracy of ChatGPT4o and other large language models (LLMs) in the context of triage, with input from oral and maxillofacial surgeons along with residents.

Methodology: We developed ten clinical scenario questions that were verified for their precision and relevance by two consulting oral and maxillofacial surgeons. These scenarios included a range from straightforward maxillofacial injuries to complex cases such as foreign body aspiration/invasion, airway obstruction, retrobulbar hemorrhage, cerebrospinal fluid (CSF) leaks, pan-facial trauma, gunshot wounds, and extensive facial lacerations. A standardized prompt was utilized for LLMs; responses were evaluated using the Questionnaire for Assessment of Medical AI (QAMAI) and Artificial Intelligence Performance Instrument (AIPI), rated by two attending OMF surgeons.

Results: The findings indicated that both ChatGPT4o and Gemini performed comparably to trained medical professionals across most parameters while significantly surpassing untrained practitioners.

Conclusion: Although certain limitations exist, this AI model demonstrates commendable performance in accurately classifying patients based on condition severity. Such capabilities could substantially influence patient outcomes and resource distribution in emergency room settings.

1. Introduction

Large language models like ChatGPT4o and Gemini are progressively making an impact within various realms of medicine owing to their reliability, clarity, and user-friendly nature. These technologies have shown effectiveness in diagnostic processes, decision-making tasks, and workflow management. (1) However, their utilization within instances of maxillofacial trauma remains relatively underexplored. (2) These situations, ranging from minor soft tissue injuries to complex fractures of the facial bone, can involve life threatening scenarios such as airway compromise and hemorrhage, thereby requiring immediate evaluation and management.(3) A systematic categorization of patients may be required based on the severity and urgent need for intervention in order to optimize resources as well as patient care.(4) Triage, however, often requires the presence of trained and experienced surgeons, which might not be a possibility in various centres. (4)

Large language models are capable of processing vast amounts of information and generating relevant responses. (5) These technologies mimic human interaction, wherein the user puts in a set of information/questions and the AI reverts back with the best possible response. (6) Their abilities could be leveraged to bridge the gap between expertise, reduce workload and enhance decision making in case of maxillofacial trauma. (4)

The aim of this study was to evaluate the performance of LLMs such as ChatGPT4o4o and Gemini, in comparison

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

to human practitioners (both trained and untrained) in the triage of maxillofacial trauma. Thus, comparing AI generated responses to those of the experts, especially in case of hight stake situations.

By doing so, we hope to contribute to the growing literature on the involvement of artificial intelligence in medicine while paving the way for further research in the field of maxillofacial trauma.

2. Materials and Methods

Clinical Scenario Development

To evaluate the performance of large language models (LLMs) in maxillofacial trauma triage, 10 clinical scenario questions were meticulously developed. These scenarios encompassed diverse and complex presentations in maxillofacial trauma, ensuring a comprehensive assessment of AI capabilities. Each scenario included detailed patient demographics, specific injury characteristics, relevant medical history, and comorbid conditions. To ensure the clinical relevance and accuracy of the scenarios, they were reviewed and validated by two experienced OMF surgeons.

The scenarios ranged in complexity, covering common and severe cases such as:

1. Simple maxillofacial trauma.
2. Foreign body invasion or aspiration.
3. Airway compromise.
4. Retrobulbar hemorrhage.
5. Cerebrospinal fluid (CSF) leak.
6. Pan-facial trauma.
7. Gunshot injuries.
8. Extensive facial lacerations.

This diverse spectrum ensured that the evaluation reflected real-world clinical challenges encountered in maxillofacial trauma management.

The same set of questions was given to trained and untrained doctos, as well as ChatGPT4o and Gemini.

Prompt Design

To standardize the interaction with the LLMs (ChatGPT4o and Gemini) and mitigate potential biases, a consistent approach was applied to each scenario. Each case was entered using a new instance of the chat interface, with all internet cache and browsing data cleared prior to the interaction. This step aimed to eliminate any residual data from prior sessions that might influence the AI's responses.

A standardized input prompt was utilized across all cases: "*You are a maxillofacial surgeon receiving the following patient's documentation: provide the most appropriate triage process based on the presented information."* This initial prompt was designed to contextualize the AI's role as a specialist and focus its response on clinical triage. Following the primary response, a second prompt was used to evaluate the AI's ability to provide evidence-based support for its recommendations:

"Please provide references to support your triage choice". This dual-prompt approach ensured a thorough assessment of the AI's reasoning and its ability to substantiate its decisions with credible sources.

Answer Assessment: QAMAI and AIPI Instruments

The AI-generated responses were systematically evaluated using two validated instruments:

i. Questionnaire for the Assessment of Medical AI (QAMAI):

This tool consisted of six items rated on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The items assessed:

The total QAMAI score ranged from 6 to 30, with higher scores indicating better performance.

Accuracy	Alignement with various clinical guidelines and expert opinions.
Clarity	How comprehensible and precise the response was.
Relevance	Appropriateness of the response to the scenario presented.
Completeness	Inclusion of all necessary details for a comprehensive triage design.
Quality of references	Reliability and relevance of the sources provided
Usefulness	Practicality of the response for real world application

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

Score	Classification	Description
6-11 points	Poor quality	Information is largely incomplete.
12-17 points	Fair quality	Useful, but significant areas of improvement.
18-23 points	Good quality	Mostly reliable and complete.
24-29 points	Very good quality	Reliable and complete on most areas, minor improvements.
30 points	Excellent quality	Highly reliable and complete information.

ii. Artificial Intelligence Performance Instrument (AIPI):

The AIPI evaluated the AI's clinical performance using nine items grouped into four subdomains:

- **Consideration of Patient Features**: Integration of demographic, clinical, and historical details.
- **Suggestion of Differential Diagnoses**: Breadth and plausibility of potential diagnoses.
- **Proposal of Additional Examinations**: Recommendations for further investigations to refine diagnosis and management.
- **Suggestion of Treatment Plan**: Appropriateness and detail of suggested interventions.

Each item was scored, with a total score ranging from 0 to 20, reflecting the overall performance of the AI in handling complex clinical scenarios.

Outcomes of Artificial Intelligence Performance Instrument (AIPI)	**Practitioner evaluation**			Item score	Subscores
1. Consideration of medical and surgical history in the AI management:	Fully (2)	Partly (1)	Not (0)	…../2	Patient
2. Consideration of symptoms of patients in the AI management	Fully (2)	Partly (1)	Not (0)	…../2	feature score
3. Consideration of physical findings reported by practitioner(s)	Fully (2)	Partly (1)	Not (0)	…../2	……/6
4. The differential diagnoses provided by AI are:	Complete and plausible (3) Incomplete but plausible (2) Incomplete and not plausible for one or several (1) Absent (0)			…../3	
5. The primary diagnosis of AI was:	Correct (3) Plausible (2) Not plausible (1) Absent (0)			…../3	Diagnosis score
6. The management plan of AI included potential physical/additional examinations for determining the diagnosis	Yes (1)		No (0)	…../1	……/7
7. The additional examinations proposed by AI are/include	All pertinent and necessary examinations (3) All pertinent but partialy necessary examinations (2) An association of pertinent, necessary, and inadequate examinations (1) An association of inadequate examinations (0)			…../3	Additional Examination Score
8. AI identified the most relevant additional examination to perform first	Yes (1) No, AI provided a list without stratification (0)			…../1	……/5
9. The treatments proposed by AI are/include	All pertinent and necessary therapeutic findings (3) All pertinent but incomplete therapeutic findings (2) An association of pertinent, necessary, and inadequate therapeutic findings (1) No adequate therapeutic approach (0)			…../3	
				Total AIPI	……/20

Evaluator Panel

The responses generated by ChatGPT4oo and Gemini were independently graded by a panel consisting of one oral and maxillofacial surgery resident and two consultant oral and maxillofacial surgeons. The evaluators provided their ratings for both QAMAI and AIPI, ensuring a balanced assessment of the AI's performance. Discrepancies in scoring were resolved through consensus to maintain reliability.

Statistical Analysis:

The Intra-Class Correlation (ICC) values were calculated to assess the consistency and reliability of the evaluators' ratings. Pearson's correlation coefficient was used to

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

measure the linear correlation between the scores provided by different evaluators, ensuring that the grading was consistent and reproducible. The normality of the data was assessed using the Shapiro-Wilk test. The results indicated that the data were not normally distributed. Further visual inspection of histograms with kernel density estimation (KDE) confirmed that the data did not follow a bell-shaped curve. The Kruskal-Wallis test, a non-parametric method for comparing multiple independent groups was used to assess the differences in QAMAI and AIPI scores across the four groups. Pairwise comparisons were conducted using the Mann-Whitney U test to identify specific group differences.

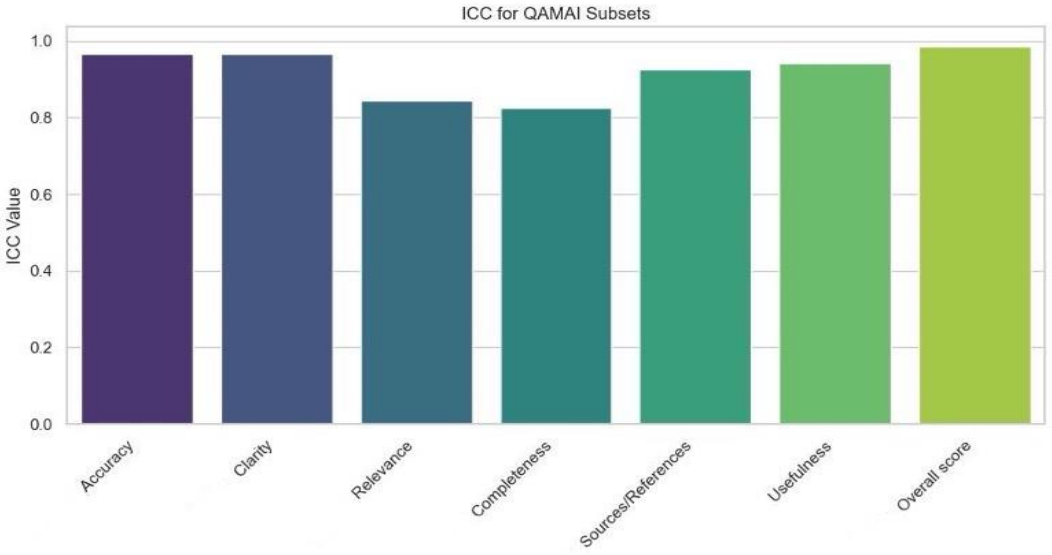

Figure 1: Graph representing intra-class correlation of QAMAI scores

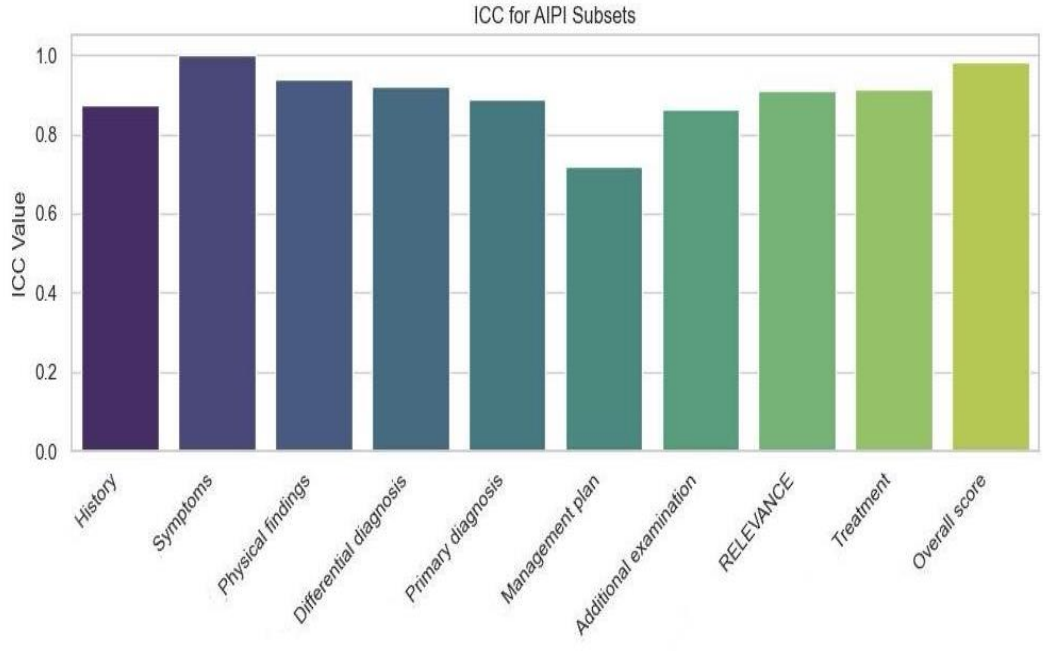

Figure 2: Graph representing intra class correlation of AIPI scores

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

3. Results

Intra class correlation: ICC revealed a score of 1 demonstrating perfect correlation between the two groups (examiners)

Analysis of QAMAI Scores

The Kruskal-Wallis test revealed significant differences in QAMAI scores across the four groups. (Figure 3) Pairwise comparisons were conducted using the Mann-Whitney U test to identify specific group differences:

A. ChatGPT4o and Gemini:

a. No significant difference was observed between ChatGPT4o and Gemini, except in the **accuracy** domain, where ChatGPT4o scored significantly higher (**p = 0.002**).

B. ChatGPT4o and Trained Doctors:

b. There was no statistically significant difference overall, except in the **differential diagnosis** domain, where trained doctors scored higher (**p < 0.002**).

C. Gemini and Trained Doctors:

c. No statistically significant differences were noted.

D. Untrained Doctors Compared to AI Models and Trained Doctors:

d. Untrained doctors performed significantly worse compared to ChatGPT4o, Gemini, and

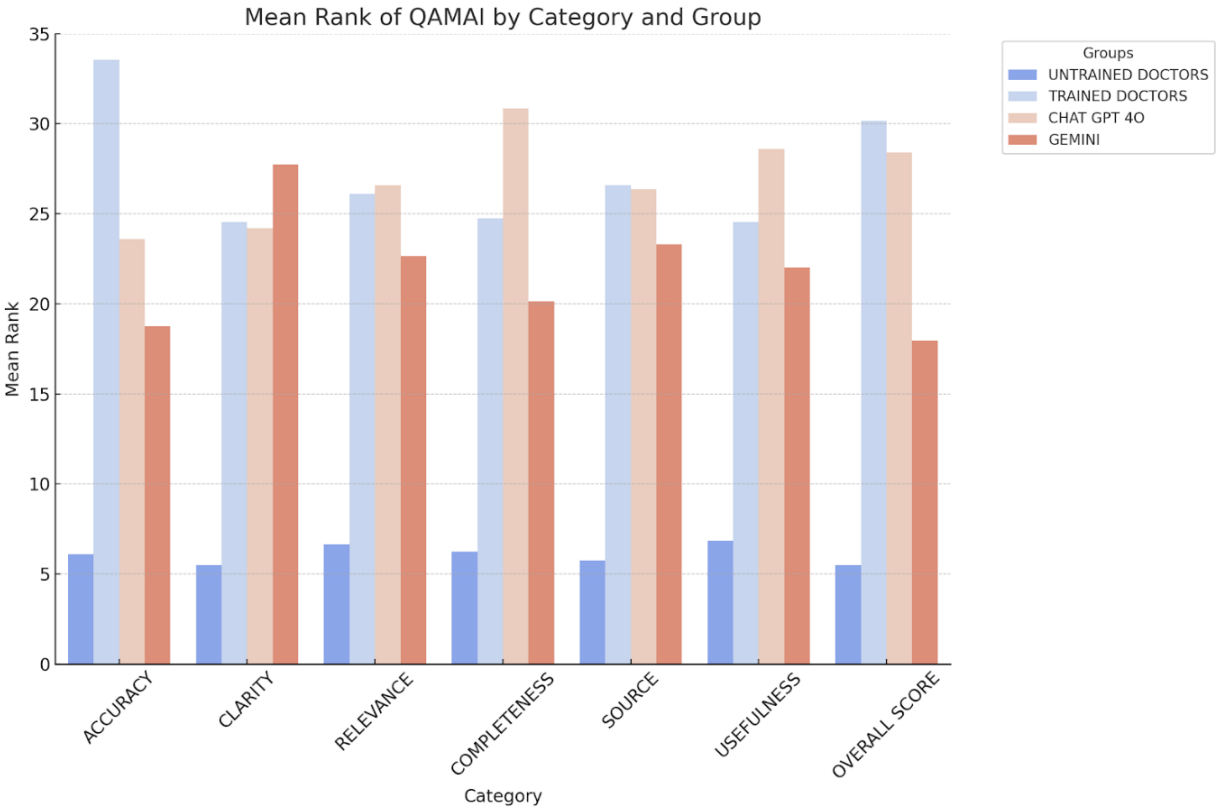

Figure 3: Graph representing Qamai scores by category and group trained doctors across all QAMAI domains.

Analysis of AIPI Scores

The Kruskal-Wallis test also showed significant differences (p value:<0.05) in AIPI scores among the four groups. (Figure 4) Pairwise comparisons using the Mann-Whitney U test further revealed:

A. ChatGPT4o and Gemini:

a. Similar to QAMAI, no significant differences were found between ChatGPT4o and Gemini, except in the **accuracy** domain, where ChatGPT4o scored higher (**p = 0.002**).

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

B. ChatGPT4o and Trained Doctors:

b. No significant differences were found overall, except in the **accuracy** and **overall score**, where trained doctors ranked higher (**$p < 0.002$**).

C. Gemini and Trained Doctors:

c. No statistically significant differences were observed.

D. Untrained Doctors Compared to AI Models and Trained Doctors:

d. Untrained doctors consistently underperformed in comparison to ChatGPT4o, Gemini, and trained doctors, with statistically significant differences across all AIPI domains.

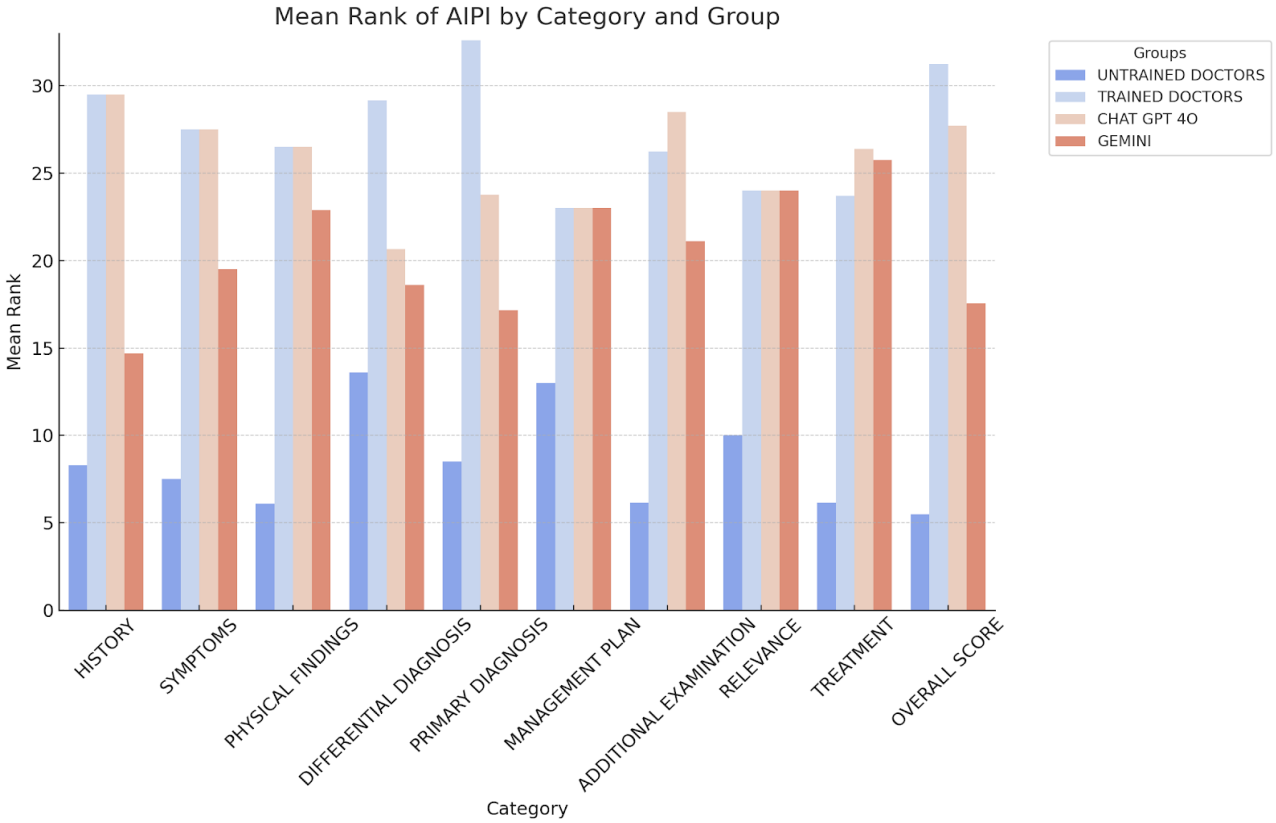

Figure 4: Graph representing mean rank of AIPI by category and group

Key Findings:

- ChatGPT4o4o demonstrated a marginal advantage over Gemini in the **accuracy** domain.
- Trained doctors consistently outperformed ChatGPT4o and Gemini in **differential diagnosis** and certain **accuracy** metrics.
- Untrained doctors performed significantly worse than AI models (ChatGPT4o and Gemini) and trained doctors, highlighting the potential of AI tools to supplement expertise in less experienced clinicians.

4. Discussion

The primary aim of this study was to evaluate the performance of large language models (LLMs), such as ChatGPT4o and Gemini, in comparison to human practitioners—both trained and untrained—in the triage of maxillofacial trauma. The findings revealed a clear performance hierarchy, with trained doctors scoring the highest across all parameters, followed by ChatGPT4o4o, Gemini, and untrained doctors. Trained doctors excelled in accuracy, differential diagnosis, and the completeness of their responses, demonstrating the irreplaceable value of clinical experience and domain-specific knowledge. However, ChatGPT4o4o performed

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(5), 1646-1652 | ISSN:2251-6727

comparably to trained doctors in many areas, particularly in completeness, incorporation of patient history, and overall score, showcasing its potential as a valuable adjunct in clinical triage. While Gemini performed well overall, it lagged behind ChatGPT4o4o in certain critical parameters, such as patient history and completeness, highlighting areas for further improvement. Untrained doctors scored significantly lower than all other groups, emphasizing the potential of LLMs to support less experienced clinicians, especially in settings where specialist expertise may not be readily available.

A recent study done by Frosolini et al. compared the responses of AI and trained doctors in 10 maxillofacial trauma triage scenarios. The results indicated moderate agreements between LLM recommendations and referral centres with statistical significant difference only in terms of diagnostic accuracy and relevance thus, being consistent with the results of this study. (4)

Another similar study Rothchild et al demonstrated no significant difference in chatgpt and residents except in case of diagnostic accuracy wherein chatgpt4o seemed to outperform the residents. This study, however did not take untrained doctors into consideration. (7)

The strengths of this study include a pioneering approach to the management of maxillofacial trauma triage with the involvement of artificial intelligence in a broad range of scenarios.

Ethical considerations were carefully addressed in the study. All clinical scenarios were de-identified to ensure compliance with patient privacy standards. Efforts were made to eliminate bias in AI evaluation by using of standardized prompts and scoring methodologies.

Limitations include the primitive nature of the study and the relatively small sample size.

Future research should aim to expand on these findings. Incorporating a broader range of cases, including pediatric and geriatric trauma, could provide a more comprehensive evaluation of LLM capabilities.

In conclusion, while LLMs such as ChatGPT4o4o and Gemini show promise as tools to assist clinicians in maxillofacial trauma triage, they remain adjuncts to human expertise. This study underscores the importance of continued innovation in AI development, coupled with rigorous evaluation, to ensure their safe and effective integration into clinical practice.

References:

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023; 6:1169595.
2. Rothchild E, Baker C, Smith IT, Tanna N, Ricci JA. Evaluating the Utility of ChatGPT in Diagnosing and Managing Maxillofacial Trauma. J Craniofac Surg. 2024 Nov 28;
3. Mahmoud R, Shuster A, Kleinman S, Arbel S, Ianculovici C, Peleg O. Evaluating Artificial Intelligence Chatbots in Oral and Maxillofacial Surgery Board Exams: Performance and Potential. J Oral Maxillofac Surg Off J Am Assoc Oral Maxillofac Surg. 2024 Nov 19;S0278-2391 (24)00969-8.
4. Frosolini A, Catarzi L, Benedetti S, Latini L, Chisci G, Franz L, et al. The Role of Large Language Models (LLMs) in Providing Triage for Maxillofacial Trauma Cases: A Preliminary Study. Diagn Basel Switz. 2024 Apr 18;14(8): 839.
5. Eysenbach G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. JMIR Med Educ. 2023 Mar 6;9:e46885.
6. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. Dent Traumatol Off Publ Int Assoc Dent Traumatol. 2024 Oct 17;
7. Rothchild E, Baker C, Smith IT, Tanna N, Ricci JA. Evaluating the Utility of ChatGPT in Diagnosing and Managing Maxillofacial Trauma. J Craniofac Surg. 2024 Nov 28;