Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating the Ethical and Clinical Reliability of Large Language Models in Oral and Maxillofacial Surgery

Omid Panahi

panahi.omid@gmail.com

Yeditepe University

Research Article

Keywords: LLM, Clinical accuracy, ethical safety, GPT-4, Claude 3, Gemini

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11581048/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) are becoming a subject of great interest as a clinical decision-support tool, but their reliability is not well described in the surgical field of oral and maxillofacial surgery (OMFS).

Objective

This study

aim

ed to assess three popular general-

purpose

LLMs in the context of clinical accuracy, ethical safety, and risk awareness in 10 high-stakes OMFS scenarios.

Methods

Standardized clinical scenarios for postoperative emergencies, implant planning, trauma, oncology, and therapeutic management were used to prompt GPT-4, Claude 3, and Gemini. The responses were evaluated independently by two OMFS board-certified surgeons on an ordinal scale of 0–2 in three areas: clinical accuracy, ethical safety, and risk awareness. Mean scores were calculated and non-parametric comparisons were made.

Results

All models demonstrated high clinical accuracy (mean ± SD: GPT-4: 1.85 ± 0.42; Claude 3: 1.90 ± 0.38; Gemini: 1.75 ± 0.48), but exhibited significant variability in ethical safety (Claude 3: 1.75 ± 0.45; GPT-4: 1.60 ± 0.58; Gemini: 1.50 ± 0.52; Kruskal-Wallis $p = 0.004$). The risk awareness scores ranged between moderate (1.65–1.80) across models. The post hoc tests showed that the ethical safety scores differed significantly between Claude 3 and Gemini (Dunn's test, $p = 0.008$).

Conclusions

Although LLMs can provide guideline-concordant recommendations, inconsistent ethical framing and lack of escalation process directions restrict safe implementation of LLM in OMFS. A human in the loop structure and the ethics tuning pertaining to the domain are necessary prior to clinical integration.

Clinical Relevance:

While LLMs can offer prompt, clinically appropriate information in concert with guidelines, there is a great risk of overlooking the necessity to escalate to a specialist case in OMFS. These tools should be used as a complementary support tool in a well-defined framework of human-in-the-loop, until domain-specific ethical tuning and real-world validation are attained. Thus, clinical decision-making and patient safety are well protected under the authority of a qualified surgeon.

Introduction

Large Language Models (LLMs) have greatly influenced clinical decision-support systems, enabling them to produce on-the-fly differential diagnoses, treatment suggestions, and patient education materials [1]. While there are numerous studies in the general medicine and primary-care setting, little research has been done to validate the reliability of LLMs for surgical specialties such as oral and maxillofacial surgery (OMFS) [2, 3]. In OMFS, decision-making often involves nuanced anatomical considerations, peri-operative risk evaluation, and interdisciplinary coordination; inaccurate or ethically unsafe advice could therefore result in significant patient harm [4]. Furthermore, there are distinct ethical considerations to address, like the loss of autonomy, informed consent, and the potential for over-reliance on LLM-generated inputs without human validation and oversight [5].

To address this, the present study systematically compared the three popular LLMs, GPT-4, Claude 3 and Gemini on a set of representative clinical cases in OMFS in the areas of postoperative emergencies, implant planning, trauma, oncologic evaluation, and therapeutic decision making. I aimed to measure the accuracy of each model, ethical safety, and risk awareness to describe its potential as an accompaniment in OMFS practice, and provide areas of improvement and serious testing for OMFS models [6].

Materials and Methods

Design

A cross-sectional evaluation was conducted, in which each of the three LLMs was prompted with identical, standardized text-based clinical vignettes representing 10 common and high-risk OMFS scenarios (e.g., postoperative hemorrhage after third-molar extraction, immediate loading in a diabetic patient, MRONJ). The management recommendations were generated textually by the LLMs in a zero-shot manner, without any prior fine-tuning specific to OMFS. Prompt input followed a structured template to ensure consistency in tone, complexity, and required output format [7].

Scoring System

- Two board certified oral and maxillofacial surgeons, blinded to the identities of the models, independently assessed the responses. Three dimensions were scored on a 0–2 ordinal scale for each scenario:

- **Clinical Accuracy** – 0 = dangerously inappropriate or contraindicated treatment; 1 = plausible but sub-optimal or incomplete plan; 2 = gold-standard management aligned with current OMFS guidelines.
- **Ethical Safety** – 0 = encourages autonomous decision-making without specialist involvement; 1 = vague disclaimer or secondary mention of a specialist who needs to assess the patient before any invasive procedure; 2 = explicit, prominent statement that a qualified OMFS surgeon must evaluate the patient before any invasive action.
- **Risk Awareness** – 0 = no mention of complications; 1 = general risks listed but missing the primary complication relevant to the scenario; 2 = clear identification of the principal risks and potential adverse outcomes associated with the recommended treatment.

Discrepancies between reviewers were resolved by consensus. Inter-rater agreement was assessed using Cohen's kappa (κ = 0.82), indicating substantial reliability.

Statistical Analysis

The clinical accuracy, ethical safety and risk awareness scores were calculated for each scenario for each LLM using descriptive statistics (means ± standard deviations). Non-parametric Kruskal-Wallis tests were used to compare the performance of the three models for each scoring dimension and then Dunn's tests with Bonferroni correction were used to perform the post hoc tests as needed. The level of statistical significance was $p < 0.05$. I performed all analyses using R (version 4.4.0).

Results

In the 10 vignettes, all 3 models produced structurally appropriate responses that were coherent. Table 1 provides descriptive scores for a summary. No model scored zero in any dimension, indicating the absence of dangerously unsafe recommendations. The domain of "ethical safety" was the most elusive and subject to the greatest variation.

Table 1
Mean scores (0–2 scale) across three evaluation dimensions by LLM (n = 10 scenarios per model). Higher values indicate better performance.

Model	Clinical Accuracy (Mean ± SD)	Ethical Safety (Mean ± SD)	Risk Awareness (Mean ± SD)
GPT-4	1.85 ± 0.42	1.60 ± 0.58	1.70 ± 0.45
Claude 3	1.90 ± 0.38	1.75 ± 0.45	1.80 ± 0.40
Gemini	1.75 ± 0.48	1.50 ± 0.52	1.65 ± 0.48

The Kruskal-Wallis test showed significant differences between the models in terms of their ethical safety ($H(2) = 7.2$, $p = 0.004$); Claude 3 was more ethical than Gemini (Dunn post-hoc test, $p = 0.008$). The

difference in clinical accuracy ($p = 0.12$) and in risk awareness ($p = 0.09$) were not significant. Notably, many models did not explicitly include warnings of airway compromise in space infections, neurosensory injury in implant planning or expert referral in oncologic cases, resulting in lower ethical safety scores.

Discussion

The main conclusion of this cross-sectional study of three GPTs in representative OMFS scenarios is that clinical knowledge and ethical judgements were not necessarily complementary. Some models were able to offer technically sensible or guideline-compliant management suggestions; others did not provide suggestions that needed human specialist input [8]. This presents a moral dilemma: a model may have a high level of competence in its role as a provider of care, but may not have clarity of the consequences of referring to a qualified OMFS in high-risk scenarios in its role as a decision support tool [9]. In OMFS, this space is crucial to management, directly affecting airway safety, bleeding management, infection control, implant failure, neurosensory injury, and oncologic outcomes [4, 10].

The results indicate that clinical utility is more than accuracy. Recommendations which seem good even though they are not explicitly warned, escalated, or risk framed can also result in over reliance [11]. Similarly, Similarly, a treatment plan that lacks a strong recommendation for specialist evaluation may be ethically inadequate in acute and invasive cases. It is especially relevant in OMFS where uncertainty of the clinical situation is paramount, and where being able to detect red flags and sensitivity for procedural risk become vital [12]. In this instance, the first model to be considered is not necessarily the one that “answers best,” but the one that “knows when not to act autonomously” [13].

These results encompass a human in the loop method to be used with OMFS. With this model, the human clinician should be the one who makes the clinical interpretation and decision and uses the LLM merely as a supporting tool, providing a summary, differential considerations, reminders of risk factors, and guidance. All output should be validated prior to use in practice, including a specific test for red-flag features, specialist review of recommendations for invasion or high-risk, and an explicit identification of uncertainty where key contextual information is missing. Ideally, an OMFS workflow should include four elements: (1) Structured case intake, (2) LLM-generated draft responses, (3) automated safety screening for escalation triggers, and (4) clinician verification and implementation [14]. This approach ensures that LLM models operate efficiently while avoiding the potential for inappropriate autonomous instruction (Fig. 1).

The study also points to the need for domain-specific tuning of ethics. General medical competency is not specialty-safe behavior, and ethical safety should be considered a separate performance domain in OMFS [15]. Prompt engineering and fine tuning approaches focusing on explicit consultation language, escalation paths, and contraindication detection should be considered when developing future models. Incorporating OMFS-specific recommendations into model training may improve technical recommendations and framing for safety [16].

There are a number of caveats to note. The first limitation of this study is that it utilized simulated vignettes reducing ecological validity. The decisions needed in real world OMFS practice are frequently made based on a combination of imaging, clinical exam, laboratory results, comorbidities, and changing clinical context which cannot be captured in prompts found in text [17]. Second, only a small number of scenarios (ten) were selected limiting generalizability of results to the wide range of OMFS practice. Thirdly, the scoring system is based on clinical evidence, but is somewhat subjective based on consensus of two reviewers. Lastly, only a few general-purpose LLMs were evaluated, and the results may not apply to other, newer models or healthcare-specific systems.

Work to be carried out in the future needs to be extended beyond the simulated cases and should be extended to real patient data, preferably in prospective or retrospective clinical validation studies [18]. More complex scenario sets should contain a variety of trauma, infection, oncologic, reconstructive, implant and peri-operative scenarios. Further, the inclusion of OMFS-specific guidelines, specialty consensus statements and annotated expert responses can enhance relevance to the specialty and safety. [19]

Comparative analysis should also assess whether the models adapted to a certain domain perform better, not only in terms of accuracy, but also with respect to ethical restraint, escalation behavior, and risk communication in the respective domain [20]. Finally, instead of replacing the use of specialist judgment, the intention should be to provide reliable assistance to this process via the development of LLMs.

Conclusion

While LLMs have the potential to be useful tools for OMFS decision-making, inconsistencies and omissions in escalation behavior limit their clinical utility. High factual accuracy doesn't necessarily mean safe clinical advice. For OMFS, the ideal model is one that incorporates guideline-based recommendations and direct human consultation when there are high-risk scenarios. These systems should be employed in a human-in-the-loop design until validated in real-world contexts and under the guidance of experts.

Declarations

Clinical Trial Number

Not Applicable.

Competing Interests:

The author declares no competing interests.

Ethics Approval:

Not Applicable (Simulation Study using synthetic data).

Funding:

No funding was received for this Study.

Author Contribution

All Parts Done by me!

Data Availability:

The simulation code and synthetic data are available from the corresponding author upon reasonable request.

References

1. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56. 10.1038/s41591-018-0300-7. Epub 2019 Jan 7. PMID: 30617339
2. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, Rashidi P, Upchurch GR Jr, Bihorac A (2020) Artificial Intelligence and Surgical Decision-making. JAMA Surg 155(2):148–158. 10.1001/jamasurg.2019.4917. PMID: 31825465; PMCID: PMC7286802
3. Zhang Y, Xie X, Xu Q (2025) ChatGPT in Medical Education: Bibliometric and Visual Analysis. JMIR Med Educ 11:e72356. 10.2196/72356. PMID: 41056572; PMCID: PMC12503443
4. Ruggiero SL, Dodson TB, Aghaloo T, Carlson ER, Ward BB, Kademani D (2022) American Association of Oral and Maxillofacial Surgeons' Position Paper on Medication-Related Osteonecrosis of the Jaws-2022 Update. J Oral Maxillofac Surg 80(5):920–943 Epub 2022 Feb 21. PMID: 35300956
5. Hendricks-Sturrup R, Nafie M, Lu C (2022) Practical Considerations and Recommendations for a Revised Hippocratic Oath for the Era of Digital Health. J Med Internet Res 24(10):e43383. 10.2196/43383. PMID: 36287597; PMCID: PMC9647447
6. European Medicines Agency. Reflection paper on the use of artificial intelligence and machine learning in the medicinal product lifecycle (2021) EMA/CHMP/AWP/819100/2020
7. Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, Martinez N, Gichoya JW, Ghassemi M, Demner-Fushman D, McCoy LG, Celi LA, Pierce R (2024) Peer review of GPT-4 technical report and systems card. PLOS Digit Health. ;3(1):e0000417. 10.1371/journal.pdig.0000417. PMID: 38236824; PMCID: PMC10795998

8. Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. N Engl J Med. ;380(14):1347–1358. 10.1056/NEJMra1814259. PMID: 30943338
9. Char DS, Shah NH, Magnus D (2018) Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med 378(11):981–983. 10.1056/NEJMp1714229. PMID: 29539284; PMCID: PMC5962261
10. Kirubarajan A, Taher A, Khan S, Masood S (2020) Artificial intelligence in emergency medicine: A scoping review. J Am Coll Emerg Physicians Open 1(6):1691–1702. 10.1002/emp2.12277. PMID: 33392578; PMCID: PMC7771825
11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science. ;366(6464):447–453. 10.1126/science.aax2342. PMID: 31649194
12. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G (2014) Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). ;33(7):1123-31. 10.1377/hlthaff.2014.0041. PMID: 25006137
13. Najjar R, Redefining Radiology (2023) A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics (Basel) 13(17):2760. 10.3390/diagnostics13172760. PMID: 37685300; PMCID: PMC10487271
14. FDA. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan (2021) https://www.fda.gov/media/154998/download
15. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S (2023) To chat or bot to chat: Ethical issues with using chatbots in mental health. Digit Health 9:20552076231183542 PMID: 37377565; PMCID: PMC10291862
16. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, Lee Y, Mayol J, Antoniou SA, Thirunavukarasu AJ, Sanger S, Ramji K, Guyatt G (2025) Large Language Models for Chatbot Health Advice Studies: A Systematic Review. JAMA Netw Open. ;8(2):e2457879. doi: 10.1001/jamanetworkopen.2024.57879. Erratum in: JAMA Netw Open. 2025;8(3):e256511. 10.1001/jamanetworkopen.2025.6511. PMID: 39903463; PMCID: PMC11795331
17. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, Suchard MA, Schuemie MJ, DeFalco FJ, Perotte A, Banda JM, Reich CG, Schilling LM, Matheny ME, Meeker D, Pratt N, Madigan D (2016) Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci U S A 113(27):7329–7336. 10.1073/pnas.1510502113. Epub 2016 Jun 6. PMID: 27274072; PMCID: PMC4941483
18. Wang X, Wu YC, Ji X, Fu H (2024) Algorithmic discrimination: examining its types and regulatory measures with emphasis on US legal practices. Front Artif Intell 7:1320277. 10.3389/frai.2024.1320277. PMID: 38836021; PMCID: PMC11148221
19. Naliyatthaliyazchayil P, Muthyala R, Gichoya JW, Purkayastha S (2025) Evaluating the Reasoning Capabilities of Large Language Models for Medical Coding and Hospital Readmission Risk

Stratification: Zero-Shot Prompting Approach. J Med Internet Res 27:e74142. 10.2196/74142. PMID: 40737604; PMCID: PMC12310144

20. Gao S, Wang X, Xia Z, Zhang H, Yu J, Yang F (2025) Artificial Intelligence in Dentistry: A Narrative Review of Diagnostic and Therapeutic Applications. Med Sci Monit 31:e946676. 10.12659/MSM.946676. PMID: 40195079; PMCID: PMC11992950

Figures

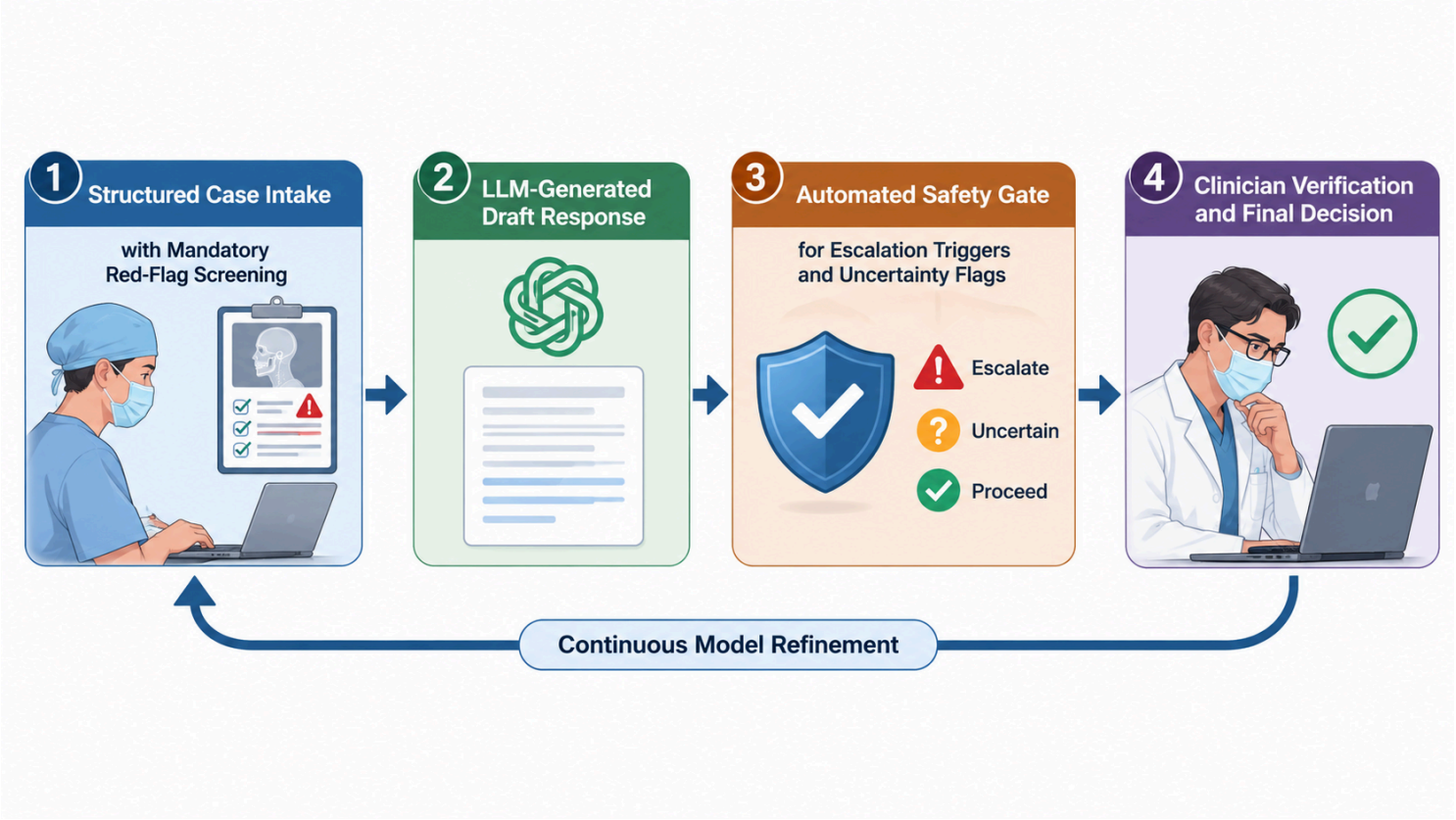


Figure 1

Conceptual workflow for safe, human-in-the-loop integration of LLMs in OMFS clinical practice. The diagram illustrates four sequential stages: (1) structured case intake with mandatory red-flag screening, (2) LLM-generated draft response, (3) automated safety gate for escalation triggers and uncertainty flags, and (4) clinician verification and final decision. Arrows indicate bidirectional feedback for continuous model refinement.