Journal of Cranio-Maxillo-Facial Surgery 53 (2025) 355–360

Contents lists available at ScienceDirect

Journal of Cranio-Maxillo-Facial Surgery

journal homepage: www.jcmfs.com

Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw

David Steybe [a,*], Philipp Poxleitner [a], Suad Aljohani [a,b], Bente Brokstad Herlofson [c], Ourania Nicolatou-Galitis [d], Vinod Patel [e], Stefano Fedele [f], Tae-Geon Kwon [g], Vittorio Fusco [h], Sarina E.C. Pichardo [i], Katharina Theresa Obermeier [a], Sven Otto [a], Alexander Rau [j], Maximilian Frederik Russe [k]

[a] Department of Oral and Maxillofacial Surgery and Facial Plastic Surgery, University Hospital, LMU Munich, Munich, Germany
[b] Department of Oral Diagnostic Sciences, Faculty of Dentistry, King Abdulaziz University, Jeddah, Saudi Arabia
[c] Department of Oral Surgery and Oral Medicine, Faculty of Dentistry, University of Oslo and Department of Otorhinolaryngology - Head and Neck Surgery Division for Head, Neck and Reconstructive Surgery, Oslo University Hospital, Oslo, Norway
[d] Oncology Patient Support Company PC, CureCancer – mycancer.gr, Athens, Greece
[e] Department of Oral Surgery, Guy's and St Thomas' Hospital, London, United Kingdom
[f] UCL Eastman Dental Institute and NIHR UCLH Biomedical Research Centre, University College London, United Kingdom
[g] Department of Oral and Maxillofacial Surgery, School of Dentistry, Kyungpook National University, Daegu, South Korea
[h] Oncology Unit, Department of Medicine and Translational Medicine Unit, DAIRI - Department of Integration, Research and Innovation, "SS Antonio e Biagio e C.Arrigo" Hospital, Alessandria, Italy
[i] Department of Oral & Maxillofacial Surgery, University Medical Center Groningen, Groningen, the Netherlands
[j] Department of Neuroradiology, University Medical Center Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[k] Department of Diagnostic and Interventional Radiology, University Medical Center Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

ARTICLE INFO

Keywords:
Large language models
GPT-4
Generative pre-trained transformer
Retrieval-augmented generation
Clinical practice guidelines
Medication-related osteonecrosis of the jaw

ABSTRACT

The potential of large language models (LLMs) in medical applications is significant, and Retrieval-augmented generation (RAG) can address the weaknesses of these models in terms of data transparency and scientific accuracy by incorporating current scientific knowledge into responses. In this study, RAG and GPT-4 by OpenAI were applied to develop GuideGPT, a context aware chatbot integrated with a knowledge database from 449 scientific publications designed to provide answers on the prevention, diagnosis, and treatment of medication-related osteonecrosis of the jaw (MRONJ). A comparison was made with a generic LLM ("PureGPT") across 30 MRONJ-related questions. Ten international experts in MRONJ evaluated the responses based on content, language, scientific explanation, and agreement using 5-point Likert scales. Statistical analysis using the Mann–Whitney *U* test showed significantly better ratings for GuideGPT than PureGPT regarding content ($p = 0.006$), scientific explanation ($p = 0.032$), and agreement ($p = 0.008$), though not for language ($p = 0.407$). Thus, this study demonstrates RAG to be a promising tool to improve response quality and reliability of LLMs by incorporating domain-specific knowledge. This approach addresses the limitations of generic chatbots and can provide traceable and up-to-date responses essential for clinical practice.

1. Introduction

Medication-related osteonecrosis of the jaw (MRONJ) constitutes a significant side effect related to antiresorptive medication, i. e. bisphosphonates and denosumab as well as other drugs like antiangiogenic medications or tyrosine kinase inhibitors (Ruggiero et al., 2022). Since its recognition as a distinct pathology more than two decades ago, MRONJ has received considerable attention in the medical community and especially in the field of oral and maxillofacial surgery. Various investigations have contributed to the understanding of this

* Corresponding author. LMU Hospital Munich, Lindwurmstr. 2a, 80337, Munich, Germany.
E-mail address: david.steybe@med.uni-muenchen.de (D. Steybe).

https://doi.org/10.1016/j.jcms.2024.12.009
Received 17 July 2024; Received in revised form 6 December 2024; Accepted 7 December 2024
Available online 10 January 2025

disease, helping in defining strategies for its prevention, diagnosis and therapy (Nicolatou-Galitis et al., 2019; Ristow et al., 2015). Recent and ongoing research continually improves and reshapes the understanding of this disease. Here, medical practice guidelines offer a way to disseminate this knowledge to clinicians. However, creating such guidelines and especially keeping them updated to new knowledge comes with a significant workload (Kredo et al., 2016).

Large language models (LLM) have made impressive progress in recent times and can address these constraints by processing, comprehending and interpreting human language. Algorithms such as ChatGPT by OpenAI are capable of passing medical and dental licensing examinations without any specialized training, indicating the potential of this technology in medical knowledge retrieval (Chau et al., 2024; Gilson et al., 2023; Kung et al., 2023). However, their applicability in clinical routine is hampered for several reasons: The data basis for the answers is neither accessible nor traceable for the user, with investigations even indicating that LLMs may generate misinformation when lacking relevant knowledge (Shen et al., 2023), a phenomenon termed *hallucination*. Additionally, access to scientific information is often restricted behind paywalls and thus not accessible for the training of LLMs like ChatGPT (Piwowar et al., 2018). Moreover, LLMs such as ChatGPT are trained with a fixed information cutoff (e. g. September 2021 in the case of GPT-4), further limiting the availability of current scientific information for the synthesis of answers (McGrath et al., 2024).

To address these limitations, knowledge-based approaches are used on retrieval augmented generation (RAG) to improve LLMs' performance (Lewis et al., 2020) and incorporating specialized medical knowledge through zero-shot learning has been demonstrated to significantly improve the performance of respective LLMs without the need for extensive retraining (Gilbert et al., 2024; Rau et al., 2023; Xiong et al., 2024; Zakka et al., 2024).

Considering the current advancements in LLMs, this study aimed to evaluate the effectiveness of a context-aware chatbot (GuideGPT) utilizing a RAG approach with scientific publications for answering questions related to the prevention, diagnosis, and treatment of MRONJ in a clinical practice guideline format. For this, we compared a context-aware LLM (GuideGPT) with a generic LLM (PureGPT) focusing on the content of the responses, language quality, scientific explanations, and the agreement of physicians with the responses.

2. Materials and methods

2.1. Technical implementation of the context-aware chatbot

The knowledge database for the context-aware chatbot was created using the references from 6 international clinical practice guidelines on MRONJ published between 2014 and 2021 (Campisi et al., 2020; Chalem et al., 2020; Kim et al., 2015; Romero-Ruiz et al., 2021; Ruggiero et al., 2014; Yarom et al., 2019). After removing duplicates and non-English references, a total of 449 articles in PDF format were obtained. The complete list of these publications is provided as Supplementary file 2. Further processing was performed using the Llama-Index software library (https://www.llamaindex.ai/). The content of the files was converted to plain text and split into individual sentences. Subsequently, each sentence was transformed into a dense vector representation using the text-embedding-ada-002 model by OpenAI, which generates high-quality sentence embeddings that encapsulate semantic meaning (Neelakantan et al., 2022). The resulting sentence embeddings were stored in a local vector store, enabling efficient similarity searches using cosine distance. When retrieving data from the vector store, a semantic similarity-based retrieval system was employed. For each query, the top 20 most semantically similar sentences were retrieved. The semantic similarity between the query and the sentences in the database was calculated using cosine similarity between their embeddings. This retrieval process ensures that the most relevant sources are prioritized in the answer generation process. Each retrieved sentence was accompanied by its surrounding context (the five sentences preceding and following it) to provide sufficient context for understanding. Additional metadata, including the file name and PDF page number from which the sentence originated, were retrieved alongside the sentences and can be presented for each answer as hyperlinks to the corresponding page of the PDF document. This functionality allows to access and verify the source material directly.

The context-aware chatbot ("GuideGPT") was implemented using GPT-4 by OpenAI, in the version of gpt-4-0613, a snapshot of GPT-4 from June 13th, 2023, as a state-of-the-art LLM with advanced natural language understanding and text generation capabilities. To ensure a balance between response consistency and creativity, we configured GPT-4 with a temperature setting of 0.4. A custom question-answering prompt template was defined to guide the chatbot's responses. The template instructs the model to provide a concise answer and explanation based on the given context, while adhering to the structure of a clinical practice guideline:

> *"Below is the scientific context information to guide your response:*
>
> *{Automatically retrieved context from PDF files}*
>
> *Based on the provided context, answer the following question:*
>
> *{Question for the Chatbot}*
>
> *Ensure your response adheres to the Clinical Practice Guidelines for MRONJ structure:*
>
> *Answer: (Provide a concise and comprehensive answer summarizing the key points.)*
>
> *Explanation: (Offer a succinct explanation for your answer. Describe relevant factors or actions and their implications.)*
>
> *Note: The context is for understanding and should not be included in your response."*

Similarly, a refine prompt template was created to allow refinement of answers based on additional context in cases in which the retrieved content exceeded the chatbot's input capacity, as GPT-4 is limited to 8192 tokens (a measure of text length). As the expected answer length was set to 1024 tokens, this mechanism triggers when the query length consisting of the question, template and sources as well as this buffer for the answer exceeded the maximum token capability of GPT-4. The template prompts the model to modify the original answer if necessary, taking into account the additional information provided by the remaining sources. In this way, the model produces fully developed answers for the initial answering step as well as all refinement steps and avoids incomplete and therefore nonsensical responses.

The full content retrieval mechanism was set up to retrieve the most relevant sentences using the vector store index, with 20 retrieved text snippets per question. The response from this chatbot was then presented alongside the metadata consisting of the name and page of the source used, which can be used as a hyperlink to access the matched content. For comparison of GuideGPT to a generic chatbot without specialized knowledge, we created a comparable setting using the same generic GPT-4 version with a similar precision prompt with the same structure and template while excluding the additional context ("PureGPT").

2.2. Evaluation of the chatbots' performance

A set of 30 questions addressing topics related to the prevention, diagnosis, and treatment of MRONJ was prepared, encompassing various aspects from preventive measures to treatment of recurrent disease (see Supplementary Table 1). Subsequently, both *PureGPT* (generic LLM) and *GuideGPT* (context-aware LLM) were tasked with responding to all of these 30 questions separately. The resulting content was then subjected to an evaluation process without any prior

alterations to the responses generated by the chatbots, except for the deletion of statements like "As of my knowledge cutoff in September 2021″ and removing the hyperlinks of the sources to maintain randomization. For the evaluation process, the sequence of the answers was randomized using MS Excel (Microsoft Corporation, Redmond, Washington, United States). This was followed by entering all answers into an online form created in MS Forms (Microsoft Corporation, Redmond, Washington, United States). Assessment was carried out independently by a panel of 10 physicians, each possessing high clinical and scientific expertise in the field of MRONJ, as evidenced by respective publications in PubMed-indexed journals (Fig. 1). For each answer and explanation provided by the LLMs, the evaluators responded to the following set of questions, applying a Likert scale ranging from 1 to 5.

- **Content:** To what extent do you agree that the content provided by the chatbot covers all relevant aspects of the question?
- **Language:** How would you rate the chatbot's use of language in terms of scientific accuracy and appropriateness?
- **Scientific explanation:** How confident are you in the scientific accuracy and evidence-based support of the explanation provided by the chatbot?
- **Agreement:** To what degree do you align with the answer and explanation provided by the chatbot?

2.3. Statistical analysis

The statistical analysis was performed using Python with the SciPy library. The median and interquartile range (IQR) were calculated for each domain (content, language, scientific explanation, agreement) for both the generic and context-aware versions of ChatGPT.

To compare the performance between the two versions, the Mann–Whitney *U* test was employed. A p-value less than 0.05 was considered statistically significant.

2.4. Code availability

All relevant code for reproducing the GuideGPT implementation is available via GitHub under the open source MIT license (https://github.com/maxrusse/GuideGPT). Use of the code for research and other projects must be in accordance with the terms of the license.

3. Results

In this study, a generic and a context-aware version of ChatGPT were tasked with answering 30 questions concerning the prevention, diagnosis, and treatment of MRONJ, resulting in a total of 60 responses. These responses were evaluated regarding 4 domains (content, language, scientific explanation and agreement) by 10 physicians possessing high clinical and scientific expertise (median of articles related to MRONJ published in PubMed indexed journals: 13) in MRONJ. Assessment was conducted on a 5-point Likert scales. All questions and responses generated by the two versions of ChatGPT, including the evaluation results for each response, are provided as supplementary table material (Supplementary Table 1).

Overall, Likert ratings ≥4 were found in 69,67% (PureGPT) and 80,67% (GuideGPT) regarding content, in 81.67% (PureGPT) and 85% (GuideGPT) regarding language, in 61% (PureGPT) and 71% (GuideGPT) regarding scientific explanation, and in 58,67% (PureGPT) and 69% (GuideGPT) regarding agreement (Fig. 2).

For the generic version of ChatGPT, statistical evaluation revealed an overall median score of 4 (IQR: 2) for content, 4 (IQR: 2) for language, 4 (IQR: 2) for scientific explanation and 4 (IQR: 1) for agreement. For the content aware version of ChatGPT, statistical evaluation revealed an overall median score of 4 (IQR: 2) for content, 4 (IQR: 1) for language, 4 (IQR: 2) for scientific explanation and 4 (IQR: 1) for language (Table 1). With the content aware version of ChatGPT surpassing the generic version in all domains (Fig. 2), these differences were statistically significant for content ($p = 0.006$) scientific explanation ($p = 0.032$), and agreement ($p = 0.008$), whereas the differences for language ($p = 0.407$) were not significantly different (Table 2).

4. Discussion

Clinical practice guidelines based on high-quality peer-reviewed research play a significant role in patient care. Traditionally, their preparation is based on manual literature search and evaluation, a task that comes with a significant workload (Kredo et al., 2016). In this present investigation, we could demonstrate the potential of combining a LLM with RAG to obtain evidence-based answers to questions related to clinical aspects of MRONJ, highlighting the potential of this technology in supporting clinical decision-making. With an already high performance of the generic version of GPT-4 by OpenAI in answering clinical questions, a context aware chatbot based on GPT-4 by OpenAI and using RAG could be demonstrated to enable significant improvement of the performance of the LLM across the domains content, scientific explanation, and agreement.

As RAG is a relatively recent development in the field of LLMs, most publications on the use of LLMs in oral and maxillofacial surgery focus on generic LLMs. In a review article published in late 2023 Puladi et al. (2024) identified three main areas of application of such LLMs in oral and maxillofacial surgery, being 1. research and scientific writing, 2. patient information and communication and 3. medical education. The general conclusion of the studies published on these aspects was that LLM like ChatGPT hold significant potential in assisting medical

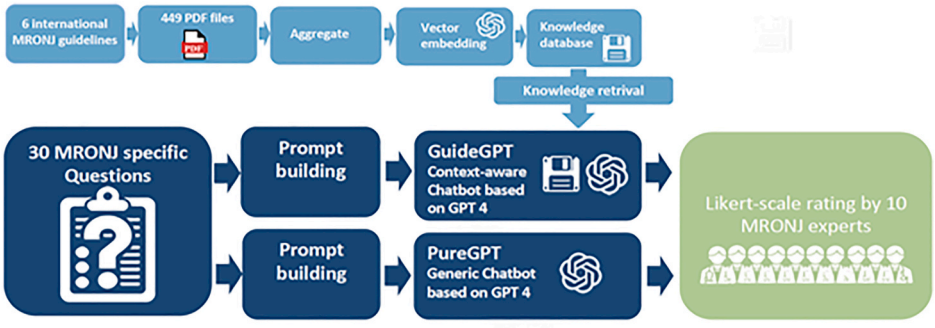

Fig. 1. Simplified representation of the technical workflow
A knowledge database was created from 449 scientific publications retrieved from 6 international MRONJ guidelines using the Llama-Index software library. A chatbot integrated with this knowledge database (GuideGPT) and a generic chatbot (PureGPT) were tasked with answering 30 MRONJ-specific questions using custom question-answering prompt templates designed to guide the chatbots' responses. Subsequent ratings were performed by 10 international MRONJ experts using 5-point Likert scales.

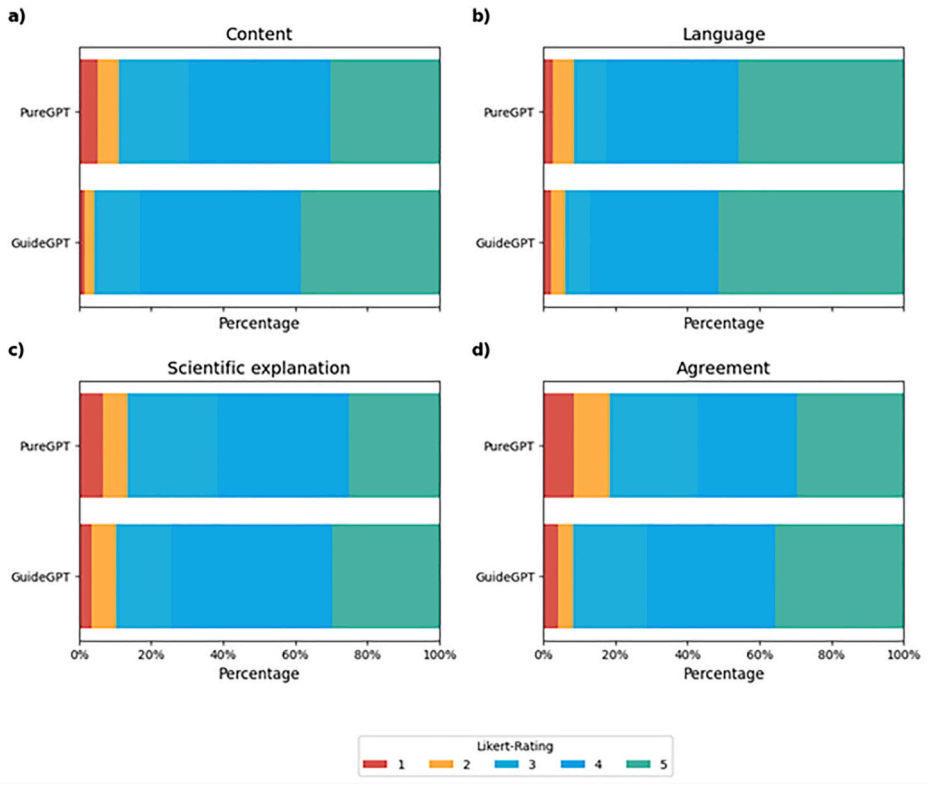

Fig. 2. Likert scale rating of the answers provided by GuideGPT and PureGPT
Bar graphs visualizing the rating results (Likert score in percent) from 10 international MRONJ experts reagrding the domains content, language, scientific explanation and agreement.

Table 1
Median and IQR for both LLMs and all domains evaluated.

	PureGPT		GuideGPT	
task/domain	median	IQR	median	IQR
Content	4	2	4	1
Language	4	1	4	1
Scientific explanation	4	2	4	2
Agreement	4	2	4	2

Table 2
Comparison of GuideGPT and PureGPT based on Likert scales.

task/domain	Statistic	p-value
Content	2.752	0.006
Language	0.83	0.407
Scientific explanation	2.141	0.032
Agreement	2.666	0.008

Statistical analysis was performed by Mann–Whitney *U* test.

research and patient information. However, the nature of current generic LLM which are trained with non-preselected publicly available data and do not provide insight into sources applied for the synthesis of answers is a major concern when applying this technology in the medical field. This is of special importance when it comes to clinical decision-making, in which traceable, reliable and up-to-date sources are indispensable.

To date, there are only few investigations reporting on the application of LLMs in the context of clinical decision making in oral and maxillofacial surgery. In a recent study, Azadi et al. (2024) investigated five different LLMs for responding to a set of 50 questions covering various aspects of OMFS and found that ChatGPT-4, among the LLMs assessed, received the highest median scoring of 4.00 with an interquartile range of 2.00 on a Likert scale ranging from 1 to 5. In another investigation conducted by Vaira et al. (2024), ChatGPT was tasked with answering 144 questions (72 open-ended, 72 binary) spanning across 12 subspecialties within head and neck surgery. When evaluated on a Likert scale ranging from 1 to 6, the overall median score for the open ended questions was 6 (interquartile range 5–6) for accuracy and 3 (IQR 2–3) for completeness. Overall, the reviewers rated the answers as entirely or nearly entirely correct in 87.2% of cases and comprehensive and covering all aspects of the question in 73% of cases. However, it was observed that 50% of bibliographic references provided by ChatGPT were nonexistent. Moreover, the authors emphasized the necessity of considering that responses provided by ChatGPT may lack current advances, discoveries or changes in medical practice and knowledge due to the fixed knowledge cutoff.

Regarding the findings of the present investigation, a closer examination of responses in which *GuideGPT* outperformed *PureGPT* revealed that in these cases, the former provided concise, and evidence-based answers, whereas the latter tended to give lengthy but vague responses. This was e. g. the case in question 2 (*"Does administration of denosumab after previous bisphosphonate therapy increase the risk of developing MRONJ?"*). Other challenges included tasks in which PureGPT's answers were not aligned with current scientific knowledge (Question 14: *"Should bone turnover markers such as CTX be determined as part of diagnosis and treatment planning in patients with suspected MRONJ?"*) or were simply incorrect (Question 26: *"Are there local/systemic adjuvant treatment options that can improve the outcome in patients undergoing surgical MRONJ therapy?"*) in which PureGPT incorrectly suggested bisphosphonates and denosumab as treatments.

However, there were also instances where *PureGPT* outperformed *GuideGPT*. For example, in Question 1 (*"Which concomitant medications (drug classes) increase the risk of MRONJ in patients under antiresorptive therapy?"*), *PureGPT* provided accepted concomitant medications along

with a clear explanation of the underlying mechanisms, whereas *GuideGPT* incorrectly stated bisphosphonates as a concomitant medication and subsequently deviated from the original question, potentially due to insufficient supporting scientific evidence in the provided literature.

As an additional and expected finding, it was observed that in questions with limited discourse in the medical community and scientific literature, both models tended to perform at a comparable level e. g. in question 13 (*"Are there any characteristic changes in the blood count of patients with MRONJ?"*).

Overall, the present investigation demonstrated significantly better results of GuideGPT in answering questions related to all clinically relevant subfields of MRONJ, regarding the domains *content* ($p = 0.006$), *scientific explanation* ($p = 0.032$) and *agreement* ($p = 0.008$). Additionally, the findings related to the domain *language* demonstrated the high language quality of answers created by LLMs such as GPT by open AI, which seems to remain unaffected by the RAG approach.

The RAG approach applied in this investigation allows simple modification to incorporate evolving scientific knowledge, as its architecture enables dynamic updating of the external knowledge source (Lewis et al., 2020). By regularly updating the knowledge retrieval component with the latest scientific literature, the model can easily be adapted to new findings and research. This flexibility makes the RAG approach an attractive option for tasks that require up-to-date information, or by incorporating domain-specific knowledge sources that allow the model to specialize in particular scientific areas. Importantly, RAG not only presents answers, but also allows users to explore topics in more depth by accessing the referenced sources via hyperlinks, a feature that is particularly valuable when encountering conflicting information or answers that require deeper understanding.

Challenges arise, however, when the included sources offer different or conflicting knowledge. In such cases, GuideGPT can explicitly acknowledge these discrepancies as part of the answer generation process. Further research is needed to address and resolve these differences during the retrieval or answer generation process. In particular, with respect to conflicts between local and global guidelines, future studies could explore methods for evaluating and highlighting relevant information through improved source text evaluation, such as re-ranking or filtering during retrieval, or guiding the chatbot based on prompts to prioritize information marked with metadata for use as a prioritized source. This moderation of sources could be based on professional societies or local institutes and would enhance the applicability of the system in different clinical settings. Building on these capabilities, models such as GuideGPT can provide easy access to current scientific evidence to support evidence-based clinical decision making, while potentially contributing to the development of future clinical guidelines.

5. Conclusion

The presented approach demonstrates that context-aware chatbots, provided they have access to sufficient scientific insights to respond accurately to questions, could serve as valuable tools for medical information retrieval. This capability holds significant potential for enhancing healthcare efficiency and supporting complex decision-making and is particularly relevant as companies begin to integrate Retrieval-Augmented Generation (RAG) options into their large language models (LLMs), making this feature widely accessible to the public.

Author contributions

David Steybe: conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing – original draft; Philipp Poxleitner: data curation, methodology, writing – review & editing; Suad Aljohani, Bente Brokstad Herlofson, Ourania Nicolatou-Galitis, Vinod Patel, Stefano Fedele, Tae-Geon Kwon, Vittorio Fusco, Sarina E.C. Pichardo: data curation, writing – review & editing, Katharina Obermeier: conceptualization, methodology, writing – review & editing; Sven Otto: Methodology, supervision, writing – review & editing; Alexander Rau: supervision, writing – original draft; Maximilian Frederik Russe: conceptualization, methodology, supervision, writing – original draft. All authors approved the final version of the manuscript.

Conflicts of interest

The authors declare that they have no conflicts of interest regarding the publication of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcms.2024.12.009.

References

Azadi, A., Gorjinejad, F., Mohammad-Rahimi, H., Tabrizi, R., Alam, M., Golkar, M., 2024. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. Oral Surg Oral Med Oral Pathol Oral Radiol 137, 587–593.

Campisi, G., Mauceri, R., Bertoldo, F., Bettini, G., Biasotto, M., Colella, G., et al., 2020. Medication-related osteonecrosis of jaws (MRONJ) prevention and diagnosis: Italian consensus update 2020. Int J Environ Res Public Health 17, 5998.

Chalem, M., Medina, A., Sarmiento, A.K., Gonzalez, D., Olarte, C., Pinilla, E., et al., 2020. Therapeutic approach and management algorithms in medication-related osteonecrosis of the jaw (MONJ): recommendations of a multidisciplinary group of experts. Arch Osteoporos 15, 101.

Chau, R.C.W., Thu, K.M., Yu, O.Y., Hsung, R.T.-C., Lo, E.C.M., Lam, W.Y.H., 2024. Performance of generative artificial intelligence in dental licensing examinations. Int Dent 74, 616–621.

Gilbert, S., Kather, J.N., Hogan, A., 2024. Augmented non-hallucinating large language models as medical information curators. NPJ Digit Med 7, 100.

Gilson, A., Safranek, C.W., Huang, T., Socrates, V., Chi, L., Taylor, R.A., et al., 2023. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9, e45312.

Kim, K.M., Rhee, Y., Kwon, Y.-D., Kwon, T.-G., Lee, J.K., Kim, D.-Y., 2015. Medication related osteonecrosis of the jaw: 2015 position statement of the Korean society for bone and mineral research and the Korean association of oral and maxillofacial surgeons. J Bone Metab 22, 151–165.

Kredo, T., Bernhardsson, S., Machingaidze, S., Young, T., Louw, Q., Ochodo, E., et al., 2016. Guide to clinical practice guidelines: the current state of play. Int. J. Qual. Health Care 28, 122–128.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., et al., 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2, e0000198.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., pp. 9459–9474

McGrath, S.P., Kozel, B.A., Gracefo, S., Sutherland, N., Danford, C.J., Walton, N., 2024. A comparative evaluation of ChatGPT 3.5 and ChatGPT 4 in responses to selected genetics questions. J Am Med Inform Assoc 128.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J.M., Tworek, J., et al., 2022. Text and code embeddings by contrastive pre-training. https://doi.org/10.48550/arXiv.2201.10005.

Nicolatou-Galitis, O., Schiødt, M., Mendes, R.A., Ripamonti, C., Hope, S., Drudge-Coates, L., et al., 2019. Medication-related osteonecrosis of the jaw: definition and best practice for prevention, diagnosis, and treatment. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 127, 117–135.

Piwowar, H., Priem, J., Larivière, V., Alperin, J.P., Matthias, L., Norlander, B., et al., 2018. The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. PeerJ 6, e4375.

Puladi, B., Gsaxner, C., Kleesiek, J., Hölzle, F., Röhrig, R., Egger, J., 2024. The impact and opportunities of large language models like ChatGPT in oral and maxillofacial surgery: a narrative review. Int. J. Oral Maxillofac. Surg. 53, 78–88.

Rau, A., Rau, S., Zoeller, D., Fink, A., Tran, H., Wilpert, C., et al., 2023. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. Radiology 308, e230970.

Ristow, O., Otto, S., Troeltzsch, M., Hohlweg-Majert, B., Pautke, C., 2015. Treatment perspectives for medication-related osteonecrosis of the jaw (MRONJ). J. Cranio-Maxillofacial Surg. 43, 290–293.

Romero-Ruiz, Mm, Romero-Serrano, M., Serrano-González, A., Serrera-Figallo, Má, Gutiérrez-Pérez, Jl, Torres-Lagares, D., 2021. Proposal for a preventive protocol for medication-related osteonecrosis of the jaw. Med Oral e314–26.

Ruggiero, S.L., Dodson, T.B., Aghaloo, T., Carlson, E.R., Ward, B.B., 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, 920–943.
Ruggiero, S.L., Dodson, T.B., Fantasia, J., Goodday, R., Aghaloo, T., Mehrotra, B., et al., 2014. American association of oral and maxillofacial surgeons position paper on medication-related osteonecrosis of the jaw—2014 update. J. Oral Maxillofac. Surg. 72, 1938–1956.
Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih, G., et al., 2023. ChatGPT and other large language models are double-edged swords. Radiology 307, e230163.
Vaira, L.A., Lechien, J.R., Abbate, V., Allevi, F., Audino, G., Beltramini, G.A., et al., 2024. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol. Head Neck Surg. 170, 1492–1503.
Xiong, G., Jin, Q., Lu, Z., Zhang, A., 2024. Benchmarking retrieval-augmented generation for medicine. https://doi.org/10.48550/arXiv.2402.13178.
Yarom, N., Shapiro, C.L., Peterson, D.E., Van Poznak, C.H., Bohlke, K., Ruggiero, S.L., et al., 2019. Medication-related osteonecrosis of the jaw: MASCC/ISOO/ASCO clinical practice guideline. J. Clin. Oncol. 37, 2270–2290.
Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., et al., 2024. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI 1. https://doi.org/10.1056/aioa2300068.