Letter to the Editor

# Reassessing the Performance of Large Language Models in Oral Health Questionnaires: Interpretative Considerations

Dear Editor,

We read with interest the article by Zhang et al, titled Comprehensiveness of Large Language Models in Patient Queries on Gingival and Endodontic Health, published in the International Dental Journal.[1] The study provides valuable insights into the potential utility of large language models (LLMs) in oral healthcare communication. However, we would like to raise several concerns regarding the study's methodology and interpretation of findings, which may have implications for the journal's readership and the broader dental community.

One notable limitation of the study is the reliance on a relatively small dataset of 33 questions. Given the vast scope of gingival and endodontic health topics, such a sample size may not sufficiently capture the variability in patient queries. The categorization of questions into 'common sense' and 'expert' groups is also somewhat ambiguous. The distinction between these categories lacks a clear operational definition, which may have introduced subjectivity in classification. Future research could benefit from a more structured taxonomy of patient queries, ensuring a broader and more representative set of questions.

Furthermore, while the study acknowledges performance discrepancies between English and Chinese responses, it does not sufficiently explore the underlying linguistic and contextual factors that may contribute to these differences. The authors attribute lower performance in Chinese to the general-purpose nature of LLMs, yet there is no discussion on how different linguistic structures,[2] character-based encoding, or disparities in available training data might have influenced these results. A deeper linguistic analysis would strengthen the study's claims and provide actionable insights for improving LLM performance in non-English medical contexts.

Another critical aspect that warrants attention is the evaluation framework used to assess the comprehensiveness of LLM responses. The study employs a five-point Likert scale; however, the specific criteria used for assigning scores remain somewhat vague. While the authors note that responses deemed inaccurate were penalized with lower scores, they do not elaborate on whether inaccuracies were factual, contextual, or interpretative. Given the clinical relevance of the subject matter, a more detailed qualitative assessment of response inaccuracies – beyond simply listing erroneous examples in supplemental materials – would enhance the reliability of the study's conclusions.

Moreover, the study's methodological approach of querying LLMs only once per question raises concerns regarding response variability. It is well-documented that LLM outputs can vary even with repeated prompts under identical conditions.[3] The decision to evaluate only a single response per model and per question does not account for this inherent variability, potentially skewing results. Repeating queries multiple times and analysing response consistency could provide a more robust assessment of LLM reliability in patient-facing applications.

Lastly, while the study acknowledges potential risks associated with misinformation in LLM-generated responses, it stops short of addressing the clinical implications of such inaccuracies. In a healthcare setting, even minor misinterpretations can lead to serious consequences, particularly when patients rely on AI-driven responses for self-diagnosis or treatment guidance.[4] The study would have been significantly strengthened by a risk assessment framework that evaluates the clinical severity of misinformation rather than just its occurrence. Additionally, the authors do not explore potential mitigation strategies, such as embedding expert verification mechanisms or integrating real-time cross-referencing with trusted medical sources.

Despite these limitations, Zhang et al contribute to an important and emerging area of research at the intersection of artificial intelligence and dentistry.[1] Their findings underscore both the promise and the current limitations of LLMs in addressing patient concerns in oral health. Future studies should aim for a more comprehensive methodological approach that accounts for linguistic complexities, response variability, and the clinical significance of AI-generated misinformation.

## Conflict of interest

None disclosed.

## Author contributions

The author contributed to the conception, analysis, interpretation of data, and drafting of the manuscript.

## REFERENCES

1. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large language models in patient queries on gingival and

endodontic health. Int Dent J 2025;75(1):151–7. doi: 10.1016/j.identj.2024.06.022.

2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. doi: 10.1038/s41586-023-06291-2.
3. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024;30(9):2613–22. doi: 10.1038/s41591-024-03097-1.
4. Al-Antari MA. Artificial intelligence for medical diagnostics-existing and future AI technology!. Diagnostics (Basel) 2023;13(4):688. doi: 10.3390/diagnostics13040688.

Carlos M. Ardila*
*Department of Basic Sciences, Biomedical Stomatology Research Group, Universidad de Antioquia U de A, Medellín, Colombia*

Pradeep Kumar Yadalam
*Department of Periodontics, Saveetha Dental College, SIMATS, Saveetha University, Chennai, Tamil Nadu, India*

*Corresponding author. Department of Basic Sciences, Biomedical Stomatology Research Group, Universidad de Antioquia U de A, Calle 70 No. 52-21, Medellín, Colombia.
E-mail address: martin.ardila@udea.edu.co (C.M. Ardila).

https://doi.org/10.1016/j.identj.2025.02.027