European Journal of Orthodontics, 2026, 48, cjag010
https://doi.org/10.1093/ejo/cjag010
Published: 8 May 2026
Reply to Letter to the Editor

Reply to Letter to the Editor

Miltiadis A. Makrygiannakis [1,2], Kostis Giannakopoulos [2], Eleftherios G. Kaklamanos [2,3,4]

[1]School of Dentistry, National and Kapodistrian University of Athens, Athens 11527, Greece
[2]School of Dentistry, European University Cyprus, Nicosia 2404, Cyprus
[3]School of Dentistry, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[4]Hamdan bin Mohammed College of Dental Medicine, Mohammed bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, P.O. Box: 505055, United Arab Emirates

Dear Prof. Eliades,

We thank the correspondents for their careful reading of our article and for their concerns, which give us the chance to further elaborate on our publication: 'Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing' [1].

Before addressing the correspondents' specific comments, we consider it important to clarify for EJO's readership the context and timing of our study, which may not have been fully taken into account in their letter. Data collection was conducted in August 2023. The article was accepted for publication in March 2024 and has been available online as an epub since April 2024. At that time, peer-reviewed orthodontic benchmarking studies in which clinical questions were posed to LLMs/LLM-powered chat systems were extremely limited, and public technical documentation regarding platform specifications, versioning, and feature configurations was comparatively scarce. Also, when the study was conducted, ChatGPT had been publicly released for <1 year, and Google Bard for <6 months, and the now-familiar pattern of frequent, incremental updates across commercial AI platforms, and the replacement or rebranding of systems over time was not yet clearly established.

We acknowledge that deployed chat systems powered by LLMs differ significantly in product features, most importantly, the availability of web-connected retrieval, and that these differences can affect how evidence-grounded a response appears. However, the purpose of our study was not to catalogue or compare the technical specifications of each platform. Rather, our objective was pragmatic: to assess the performance of publicly accessible, LLM-powered chatbot tools as they are encountered by clinicians seeking rapid, single-query support, rather than to isolate 'base-model reasoning' under feature-matched conditions. This approach is consistent with our Methods, which reflect a one-off, clinician-like interaction (single prompt per question, without follow-up, rephrasing, or iterative refinement). Differences across systems in architecture, training data, and functional capabilities are explicitly acknowledged in the Discussion, including the statement that 'Microsoft Bing AI utilizes various learning models, such as GPT-4'.

Regarding temporal documentation, as mentioned above, we clarify that all questions were submitted to the evaluated platforms in August 2023, and the responses were re-scored four weeks later (September 2023) as part of the planned test–retest assessment. At the time of study conduct, the extent and frequency of subsequent platform updates, and the likelihood of model replacement or rebranding, were not yet as evident as they are today. We acknowledge that providing explicit query dates could improve transparency and reproducibility reporting, and we are therefore pleased to state them here. Given the fact that now, almost 2 years after the publication, all of them have been replaced; we are not sure how the authors of the letter would benefit from this information.

We agree that limiting the evaluation to 10 questions is a constraint of the present study. However, item selection was inherently bounded by the requirement that each question could be benchmarked against a robust evidence base: namely, consensus statements issued by scientific organizations or professional bodies and systematic reviews identified through searches of medical libraries and the PubMed database in high-impact, peer-reviewed journals. A larger item bank, together with repeated generations per prompt, would likely improve precision and would enable explicit estimation of response variability. Nevertheless, our design was intentionally aligned with a common real-world use case: clinicians posing single, standalone questions to obtain rapid guidance, without iterative reprompting, refinement, or sampling multiple outputs.

We also think that it is once again important to note the temporal context of the work. The questions were posed and the study conducted in 2023, with online publication in 2024, prior to the publication of the specific methodological discussions cited by the correspondents -namely, concerns about wide output dispersion from identical prompts [2] and more recent clinical safety assessment frameworks emphasizing iterative sampling to quantify hallucination rates [3]. While these later contributions provide useful direction for future benchmarking studies, they were not available to inform our original protocol.

Regarding the potential for subjectivity inherent to rubric-based grading, this concern is precisely why we implemented blinding, two independent expert raters, and formal reliability analyses. The correspondents' assertion that inter-rater reliability was not reported is therefore incorrect. Specifically, we reported that responses were anonymized and coded so that evaluators were unaware of which system generated each answer; we quantified agreement using inter-evaluator

correlations (Pearson's r and Spearman's ρ); and we reported reliability metrics, including Cronbach's alpha and the intraclass correlation coefficient (ICC) across scoring occasions and pooled scores. We additionally found no overall statistically significant differences between evaluators' scores, which supports the appropriateness of averaging ratings across evaluators. Finally, we note that the reference [4] cited by the correspondents in support of their calibration/reliability critique was published after our article became available online, and thus could not have informed the study protocol at the time the work was conducted.

In summary, we concur that cross-platform comparisons must be interpreted in light of product-level differences (including retrieval augmentation) and time-dependent model updates. Given the limitations elaborated in the manuscript, we think that our conclusions were reserved, and, for sure, we clarified the need for clinicians to be aware of the limitations of LLMs/LLM-powered chatbots.

Finally, we commend the correspondents for identifying the typographical error in the Results narrative stating that 'ChatGPT-4 answers were scored as the best… followed by… Microsoft Bing Chat' with respect to the ordering of the highest-scoring system. Importantly, the findings are presented correctly throughout the remainder of the abstract, the main text, and the tables, all of which consistently indicate that Microsoft Bing Chat achieved the highest mean score. However, we do not agree with the correspondents' suggestion that this isolated error reflects shortcomings in the rigor of the peer-review process. The European Journal of Orthodontics maintains high editorial and review standards, and we consider this an isolated typographical oversight that does not affect the underlying analyses or the study's conclusions.

Data availability

The data are available in the original article and in its online supplementary material.

References

1. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. *Eur J Orthod* 2025;**48**:cjae017. https://doi.org/10.1093/ejo/cjae017
2. Suh CH, Yi J, Shim WH *et al.* Insufficient transparency in stochasticity reporting in large language model studies for medical applications in leading medical journals. *Korean J Radiol* 2024;**25**:1029–31. https://doi.org/10.3348/kjr.2024.0788
3. Asgari E, Montaña-Brown N, Dubois M *et al.* A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digit Med* 2025;**8**:274. https://doi.org/10.1038/s41746-025-01670-7
4. Savage T, Wang J, Gallo R *et al.* Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *J Am Med Inform Assoc* 2025;**32**:139–49. https://doi.org/10.1093/jamia/ocae254