

WILL CHATGPT SOON REPLACE OTOLARYNGOLOGISTS?

Quentin Mat^{1,2}, Giovanni Briganti^{2,3,4}, Antonino Maniaci⁵, Christophe Lelubre^{2,6}

Quentin Mat

quentin.mat@humani.be

1. Department of Otorhinolaryngology, C.H.U. Charleroi, Chaussée de Bruxelles 140, 6042 Charleroi, Belgium
2. Faculty of Medicine and Pharmacy, University of Mons (UMons), Mons, Belgium
3. Department of Clinical Science, Faculty of Medicine, University of Liège, Quartier Hôpital, Avenue Hippocrate 13, 4000 Liège, Belgium
4. Faculty of Medicine, Université Libre de Bruxelles, Route de Lennik 808, 1070 Brussels, Belgium
5. Faculty of Medicine and Surgery, University of Enna "Kore", Enna, Italy
6. Department of Internal Medicine, C.H.U. Charleroi, Charleroi, Belgium

Keywords : Artificial · GPT · ChatGPT · Large language model · Otolaryngology · Medicine

Dear Editor,

The past year was marked by the fulgurating success of large language models (LLM) [1, 2]. These artificial intelligence (AI) models allow to process and generate text from a neural network architecture similar to human brain organization [1, 3]. Generative pretrained transformers (GPT) are currently among the most popular LLMs [2]. The most famous implementations, OpenAI's ChatGPT-3.5 and 4, have profoundly impacted many sectors including medicine [1–3]. GPT-4 for instance, now exceeds the passing score for the United States Medical Licensing Examinations (USMLE) [2].

Otolaryngology is one of the medical fields for which the use of ChatGPT has been most studied in clinical decision-making and significant agreement between practitioners and ChatGPT has already been shown [3, 4]. These results have alarmed many physicians, that have asked the question of whether ChatGPT could soon replace physicians [2]. We think that AI could never replace the care a specialized physician could bring to patients. Moreover, to date, different weaknesses of this AI can be highlighted limiting its impact on clinical medicine. In published studies, the clinical history was encoded by physicians, involving the use of medical terms [3, 4]. In a real setting, if patients were to provide their symptoms by themselves, various approximate words would be encoded, and it would also be more difficult for some to select the most appropriate complaints. The limited time of a medical consultation, the orientation of the medical history by the physician and the different environment are some elements that could influence the information collected. Yet, prompting is a major factor determining the quality of the output, and providing inappropriate information may lead to unnecessary or even harmful examinations and treatments [1, 2, 5]. One should also bear in mind that LLMs such as ChatGPT have been trained on large amounts of texts generally from the Internet, so that the veracity of the responses may not be guaranteed [1–3]. Indeed, GPT has not been specifically designed for

medical purposes. Its goal is to accomplish many tasks in different areas [5]. The use of a LLM with a more pronounced medical background (such as Med-PaLM 2) may improve the quality of the output [2]. In addition, most GPT are not dynamically updated, preventing some patients from accessing to recent examinations or treatments [1, 2]. For instance, the database underlying ChatGPT-3.5 (the most used currently due to its free nature) ends in September 2021 [1, 2]. The fact that most GPTs only use open-source data is another barrier to the quality of the information collected [1, 2, 5]. Also, LLMs can hallucinate [1–3], meaning it can report false information such as false references or report logical contradictions [3]. Indeed, its operations are not based on understanding the meaning of the input and output but on associative relationships between words [1–3]. Although improved by fine-tuning and reinforcement learning from human feedback (RLHF), some of these hallucinations persist in GPT-4 and partly depend on hyperparameter settings [1, 2]. Finally, concerns have been raised on the possibility of leaks of private medical data or medical error [1–3].

These considerations are however rather reassuring regarding the use of ChatGPT in otolaryngology and other medical disciplines. In summary, LLMs should be considered as a tool allowing physicians to better manage their patients and accelerate some time-consuming works (e.g. writing discharge summaries for patients) [5, 6], but will not serve as a completely self-contained substitute. A legal framework is also required for this new clinical practice [6].

Data availability Not applicable.

Declarations

Conflict of interest The authors have no conflicts of interest to disclose.

References

1. Clusmann J, Kolbinger FR, Muti HS et al (2023) The future landscape of large language models in medicine. *Commun Med (Lond)* 3(1):141. <https://doi.org/10.1038/s43856-023-00370-1>. (Published 2023 Oct 10)
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. *Nat Med* 29(8):1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
3. Briganti G (2023) How ChatGPT works: a mini review. *Eur Arch Otorhinolaryngol*. <https://doi.org/10.1007/s00405-023-08337-7>. (Published online November 22, 2023)
4. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S (2023) Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otorhinolaryngol*. <https://doi.org/10.1007/s00405-023-08326-w>. (Published online November 22, 2023)
5. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 388(13):1233–1239. https://doi.org/10.1056/NEJMs_r2214_184
6. Ethics and governance of artificial intelligence for health (2024) Guidance on large multi-modal models. World Health Organization, Geneva