

CORRESPONDENCE

Open Access



Large language models and the perils of their hallucinations

Razvan Azamfirei^{1*}, Sapna R. Kudchadkar^{1,2} and James Fackler^{1,2}

To the Editor,

We read with great interest the paper by Salvagno et al. [1] As they masterfully stated, “ChatGPT work should not be used as a replacement for human judgment, and the output should always be reviewed by experts before being used in any critical decision-making or application.” As is often the case in critical care, new technologies and apparent breakthroughs are often touted as game-changers. However, the truth usually emerges the next day when the confetti has settled, and we have to clean up the sticky mess left by gallons of printed ink mixed with our hopeful wishes.

Salvagno et al. present a ChatGPT-generated summary of three studies. As they noted, the summary was believable, albeit generic and sparse in the details. The glaring problem is that it’s completely fabricated. ChatGPT cannot access the internet, and its training dataset stops in September 2021; it has no reference to any studies published in 2023 [2]. In fact, one of the trials included in the summary, Belohlavek et al. [3], showed no improvement in functional neurological outcomes, contradicting ChatGPT’s summary.

We must understand one particular aspect of large language models, which is gracefully termed as

“hallucinations”, though “fabricating information” may be more accurate [4]. In the case of the ChatGPT summary, we are only reading a generic summary of an average study with a given title.

The following is an excerpt from a “self-reflective” ChatGPT conversation.

“Q: Briefly, what is a GPT?”

A: It is a type of deep learning algorithm that is designed to generate human-like text. [...] Once trained, GPT models can be used to generate new text that is coherent and follows the patterns and styles of the original training data.

Q: Would it be fair, then, to say that GPT models, such as yourself, use a probabilistic approach to generate text that looks like previous text that was written?

A: Yes, that is a fair statement. [...] When given a prompt, the model generates text by predicting the most likely word or phrase to follow based on the patterns it has learned from the training data.

Q: When I ask you to generate new text, are you coming up with the ideas yourself or just repackaging previously written works?

A: [...] I do not come up with the ideas myself. Instead, I generate text based on patterns and relationships within my training data. This means that I am essentially repackaging previously written works, based on what I believe to be the most probable continuation of the given input text.

This comment refers to the article available online at <https://doi.org/10.1186/s13054-023-04380-2>.

*Correspondence:

Razvan Azamfirei
razvan@jhmi.edu

¹ Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

² Department of Pediatrics, Johns Hopkins University School of Medicine, Baltimore, MD, USA



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Q: Is there any filter of correctness applied to the output?

A: As an AI language model, I do not have a built-in filter to determine the correctness of my output.”

When the model had no answer, we received whatever looked like the most probable summary of the study—a hallucination.

Of course, we will improve our ability to integrate real-time information and reduce the rate of “hallucinations”, just as the nascent field of prompt engineering evolves. Running the exact prompt used by the authors now results in a response recognizing the knowledge gap. The broader point is less about whether we should use ChatGPT, and more about whether we are using it with a full understanding of its strengths and limitations [5].

We commend the authors for their exploration of ChatGPT and some associated important ethical issues. Foremost, however, it is important to reiterate that because Chat GPT now has no access to the article it was asked to interpret, it was given an impossible task. Our goal is simply to emphasize that, whether it’s a new language model, an innovative monitoring technology, or a novel biomarker, we must be aware of our tools’ limitations. We hope that as these technologies evolve, they respond, as did Robot Model B-9 in the 1960’s television show *Lost in Space*, with “that does not compute” before spewing what it must know is a hallucination.

We offer this analogy as a conclusion. Imagine a self-driving system trained to safely navigate a car on public roadways; would we place the same system in a rocket, asking it to navigate us to low earth orbit? Likely not. The tasks seem similar, navigating, but are completely different. We can build systems to take us to Earth’s orbit, just how we’ll build systems to accurately summarize scientific articles. Our only hope is that we know whether our rocket is taking us to Kansas or the International Space Station *before* strapping ourselves to it. To again quote Robot Model B-9, “Danger”.

Acknowledgements

ChatGPT was used in the writing of this manuscript. All content generated by ChatGPT is marked as such.

Author contributions

The manuscript was drafted by RA with contributions from JF and SRK. All authors contributed to the review and editing of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JF is a Board Member of Machine Learning for Healthcare and a founder of Rubicon Health LLC. The remaining authors do not have any potential competing interests.

Received: 27 February 2023 Accepted: 4 March 2023

Published online: 21 March 2023

References

1. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. 2023;27(1):75. <https://doi.org/10.1186/s13054-023-04380-2>.
2. OpenAI. ChatGPT General FAQ. 2023. <https://web.archive.org/web/20230226171118/https://help.openai.com/en/articles/6783457-chatgpt-general-faq>. Accessed on 26 Feb 2023.
3. Belohlavek J, Smalcova J, Rob D, Franek O, Smid O, Pokorna M, et al. Effect of intra-arrest transport, extracorporeal cardiopulmonary resuscitation, and immediate invasive assessment and treatment on functional neurologic outcome in refractory out-of-hospital cardiac arrest: a randomized clinical trial. *JAMA*. 2022;327(8):737–47. <https://doi.org/10.1001/jama.2022.1025>.
4. OpenAI. ChatGPT: Optimizing Language Models for Dialogue. 2022. <https://web.archive.org/web/20230225211409/https://openai.com/blog/chatgpt/>. Accessed on 26 Feb 2023.
5. Zamfirescu-Pereira J, Wong R, Hartmann B, Yang Q, editors. Why Johnny can’t prompt: how non-AI experts try (and fail) to design LLM prompts. In: Proceedings of the 2023 CHI conference on human factors in computing systems (CHI ’23); 2023; Hamburg, Germany. <https://doi.org/10.1145/3544548.3581388>. https://www.researchgate.net/publication/368577310_Why_Johnny_Can't_Prompt_How_Non-AI_Experts_Try_and_Fail_to_Design_LLM_Prompts. Accessed on 02/26/2023.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

