pagepress

GLOBAL CARDIOLOGY
Global Cardiology 2025; **3**: 89-95
DOI: 10.4081/cardio.2025.74

EDITORIAL

Large language models to write scientific manuscripts: to be considered but not trusted

Christian Basile,[1] Stefan D. Anker,[2] Gianluigi Savarese[1]

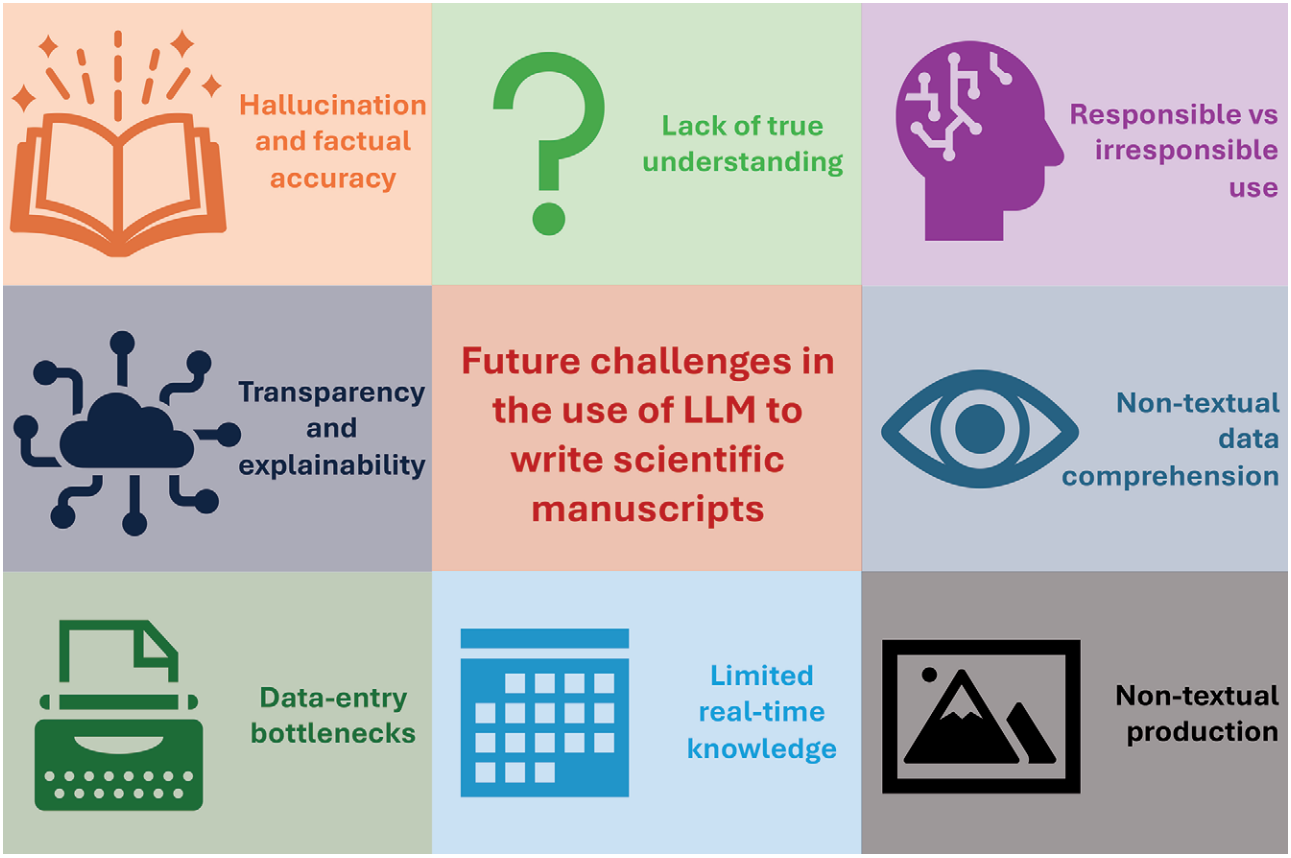
[1]Department of Clinical Science and Education, Södersjukhuset; Karolinska Institutet, Stockholm, Sweden; [2]Department of Cardiology (CVK) of German Heart Center Charité; German Centre for Cardiovascular Research (DZHK) partner site Berlin, Charité Universitätsmedizin, Berlin, Germany

Abstract

Large language models (LLMs) are rapidly permeating scientific writing. Using ChatGPT 4o and real-world data from the Swedish Heart Failure Registry, we explored both the promise and the pitfalls of LLM-assisted manuscript drafting. We demonstrate that, with strict human oversight, LLMs may accelerate text generation but still hallucinate, frequently misinterpret effect directions, and are often unable to access other manuscripts for direct references. Until these limitations are addressed, LLMs should complement, rather than replace, domain experts in scientific publishing.

Graphical abstract

Future challenges in the use of large language models to write scientific manuscripts. LLM, large language model.



Key words: artificial intelligence; machine learning; large language models; writing tools; heart failure.

Received: 28 May 2025; Accepted: 23 June 2025.
**Correspondence to: Gianluigi Savarese, MD PhD, Department of Clinical Science and Education, Karolinska Institutet; Södersjukhuset, Sjukhusbacken 10, 118 83 Stockholm, Sweden. E-mail: gianluigi.savarese@ki.se*

Introduction

Scientific publishing is essential for the advancement of science. It disseminates research findings, encourages collaboration, promotes reproducibility, and ensures that scientific knowledge is accessible and verifiable. There is currently much speculation about the widespread use of large language models (LLMs) such as ChatGPT in academic writing and how these tools might impact global scientific practices.

The surge of terms commonly used by LLM is a clear indication of the constant and increasing use of these models in writing scientific manuscripts (Figure 1).[1]

The question of whether LLMs can independently write entire manuscripts is still under debate. To investigate, we tested this hypothesis using ChatGPT 4o, an AI-powered conversational agent based on the Generative Pre-trained Transformer (GPT) LLM. We supplied ChatGPT with aggregated data from the Swedish Heart Failure Registry (SwedeHF) to examine sex differences in baseline characteristics, prescription patterns, and clinical outcomes in two distinct populations: one with an ejection fraction (EF) <40% and the other with an EF ≥40%.

Data entry

Data entry may be carried out through two primary methods. Data may either directly be copied and pasted as text into ChatGPT or uploaded as a file in PDF, Excel, or comma-separated values format. It is important to note that ChatGPT has limited ability to process PDF files, particularly those from published manuscripts, which often have complex formatting, including images and tables, with ChatGPT being also unable to analyze images or tables within a PDF file. Additionally, while Excel or comma-separated values files are visible once uploaded in ChatGPT, the LLM was only able to access them three times out of ten (Figure 2). Given these limitations, data entry was all done by copy-pasting the data inside ChatGPT (Figure 3).

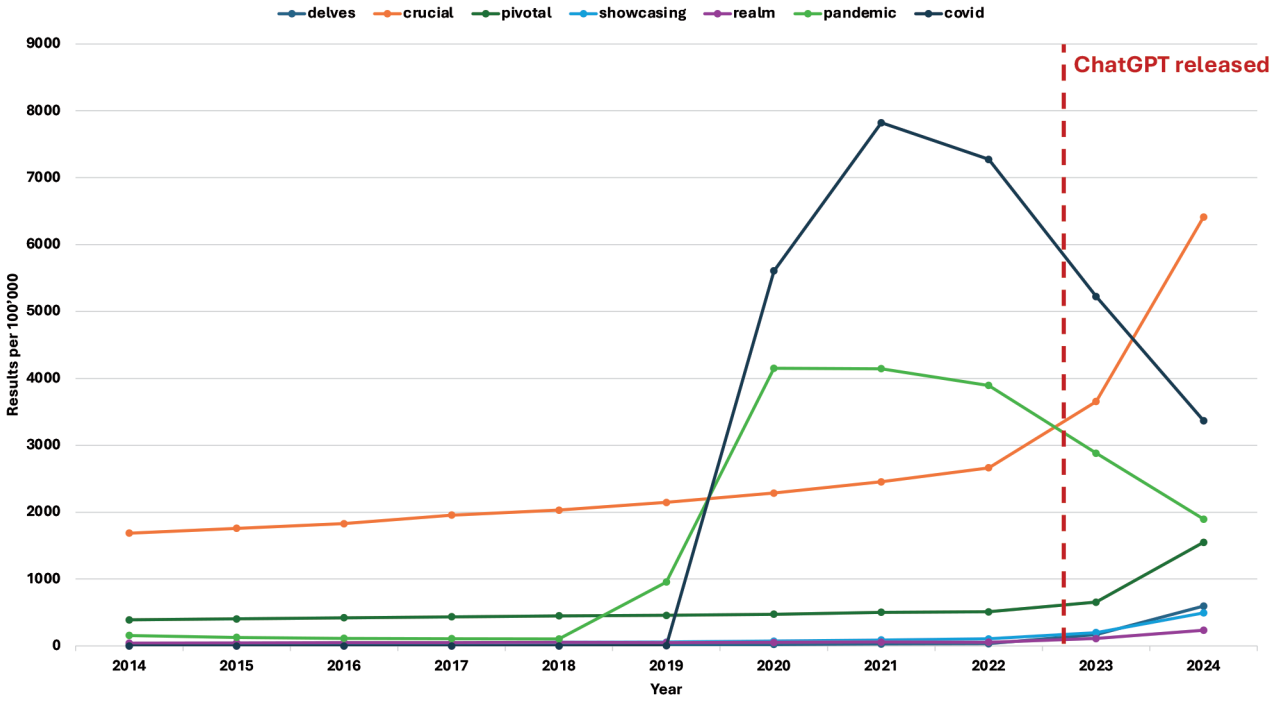


Figure 1. Word Frequency Shift in PubMed abstracts over 10 years (2014-2024). The plot shows the frequency over time for the top 5 words most disproportionately used by LLM compared with humans, together with two major events influencing scientific writing (pandemic, covid), as measured by frequency per 100.000 words. These terms maintained a consistently low frequency in PubMed abstracts until 2022 but experienced a sudden surge in usage starting in 2023 (ChatGPT released in November 2022).

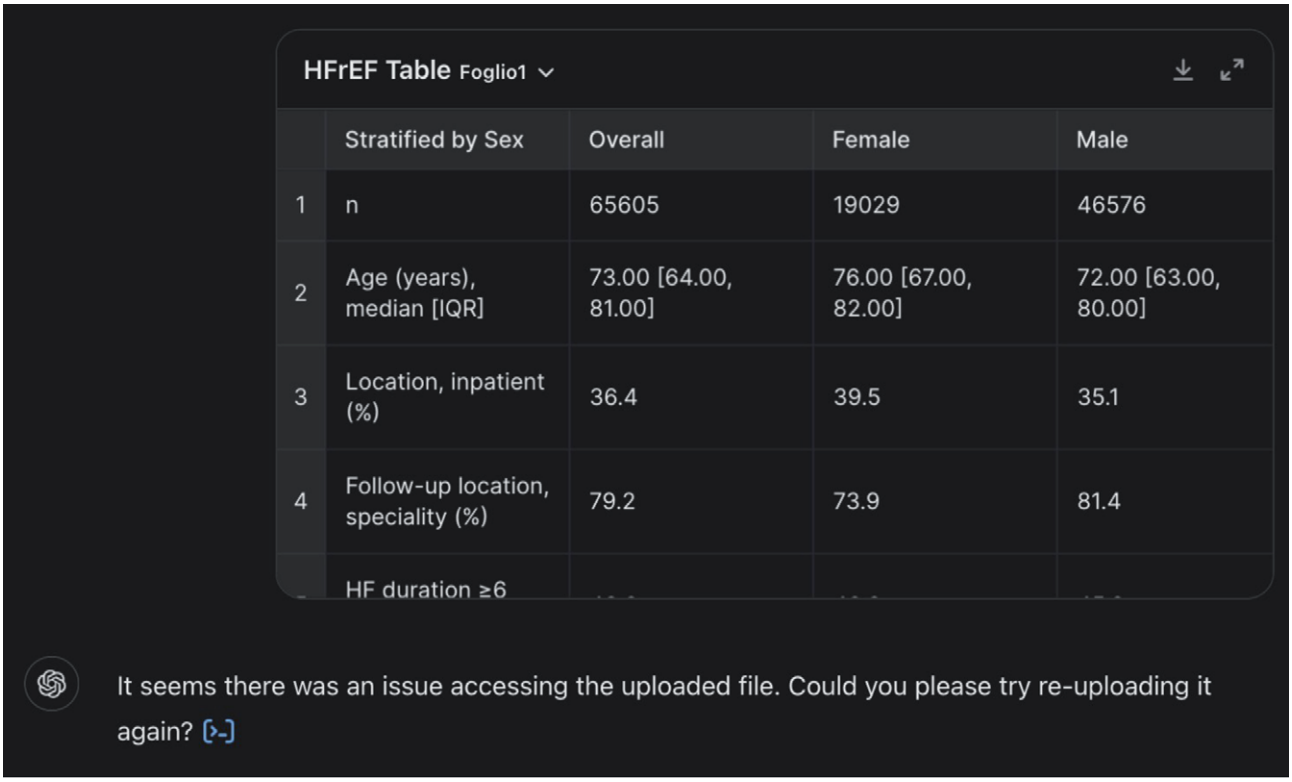


Figure 2. Example of data entry failure. Despite the table for the baseline characteristics of patients with heart failure with reduced ejection fraction being visible in ChatGPT, the LLM was not able to access and describe the uploaded Excel file.

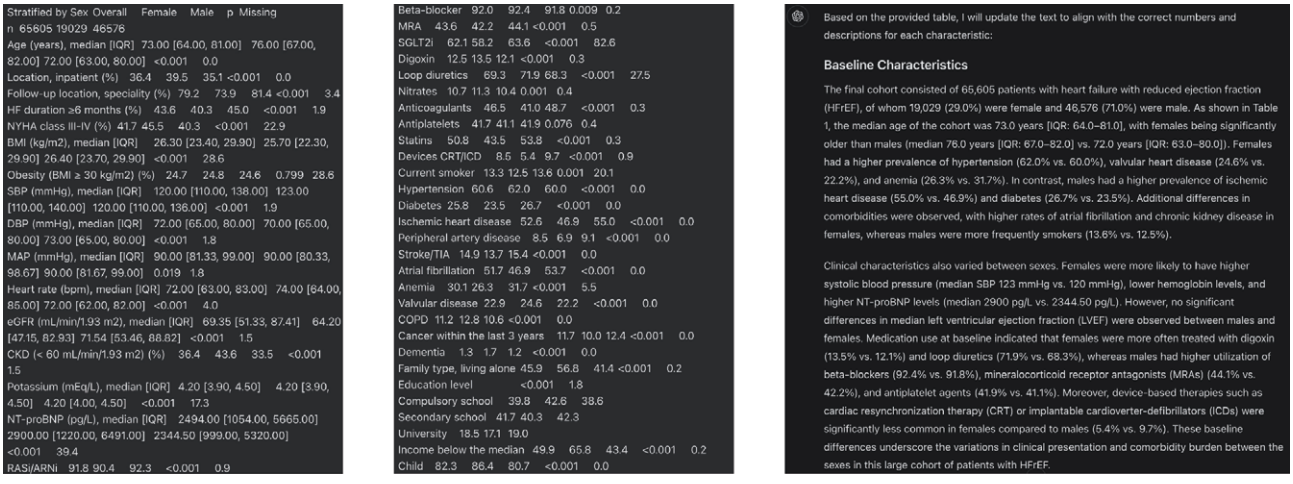


Figure 3. Example of how the data entry for aggregate data was performed (on the left and center) and how the LLM interpreted the data (on the right). A hallucination appeared clearly in the description of the median left ventricular ejection fraction, which was never supplemented as data.

Limitations

Hallucinations

In the context of LLMs, hallucination refers to generating outputs that appear credible and plausible but are factually incorrect, misleading, or entirely fabricated.[2,3] This phenomenon occurs because LLMs do not inherently distinguish between true and false information; instead, they generate text based on patterns and statistical probabilities derived from their training data. Hallucinations can be caused by several factors, including incomplete training data, ambiguous prompts, or contexts without consensus, making it difficult for the model to provide accurate answers.[4]

Hallucinations can have severe consequences on the accuracy of the prompted result. For example, a study reported that hallucination rates can range from 2.5% to 15% in common LLMs like GPT-4, Llama, and Gemini and can reach up to 88% in high-stakes legal contexts.[5] This makes hallucinations a significant challenge, as they could mislead users into making incorrect decisions based on unreliable outputs.

Despite our experiment, which was based on three simple tables per manuscript, hallucinations frequently occurred in the form of wrong/fabricated data (Figure 4). To address this problem, the data produced by the model in the text was reviewed by the authors, who compared every generated numerical value with the SwedeHF tables. Any unsupported claim was returned to the model with a corrective prompt and rechecked for accuracy.

Lack of understanding

The lack of understanding in LLMs is a fundamental challenge rooted in how these models are built and trained. Although they produce human-like responses that often seem insightful or accurate, their outputs are based purely on pattern recognition and statistical associations, not genuine comprehension of language or concepts. LLMs like GPT-3 and GPT-4

Figure 4. Examples of hallucinations with wrong (male/female sex proportions are entirely different from the one provided to the LLM) or fabricated (mean ejection fraction) data.

generate text by predicting the most probable word sequences based on the context provided by the user and the vast volumes of text they were trained on. This approach relies on surface-level coherence rather than deeper semantic understanding.[6]

One of the main reasons for this limitation is that LLMs are trained to optimize word prediction, which means they lack mechanisms to grasp abstract ideas. They do not have access to real-world experiences or any form of grounding that could enable them to differentiate between plausible and factual statements. Instead, they depend on learned associations from text data, which means they may produce outputs that sound convincing but are inaccurate or logically flawed.[7]

In our experiment, the LLM was unable to understand the directionality of odds ratios and hazard ratios despite it being clearly stated in the table. Furthermore, despite a different direction in the effect, the LLM gave the same understanding (Figure 5).

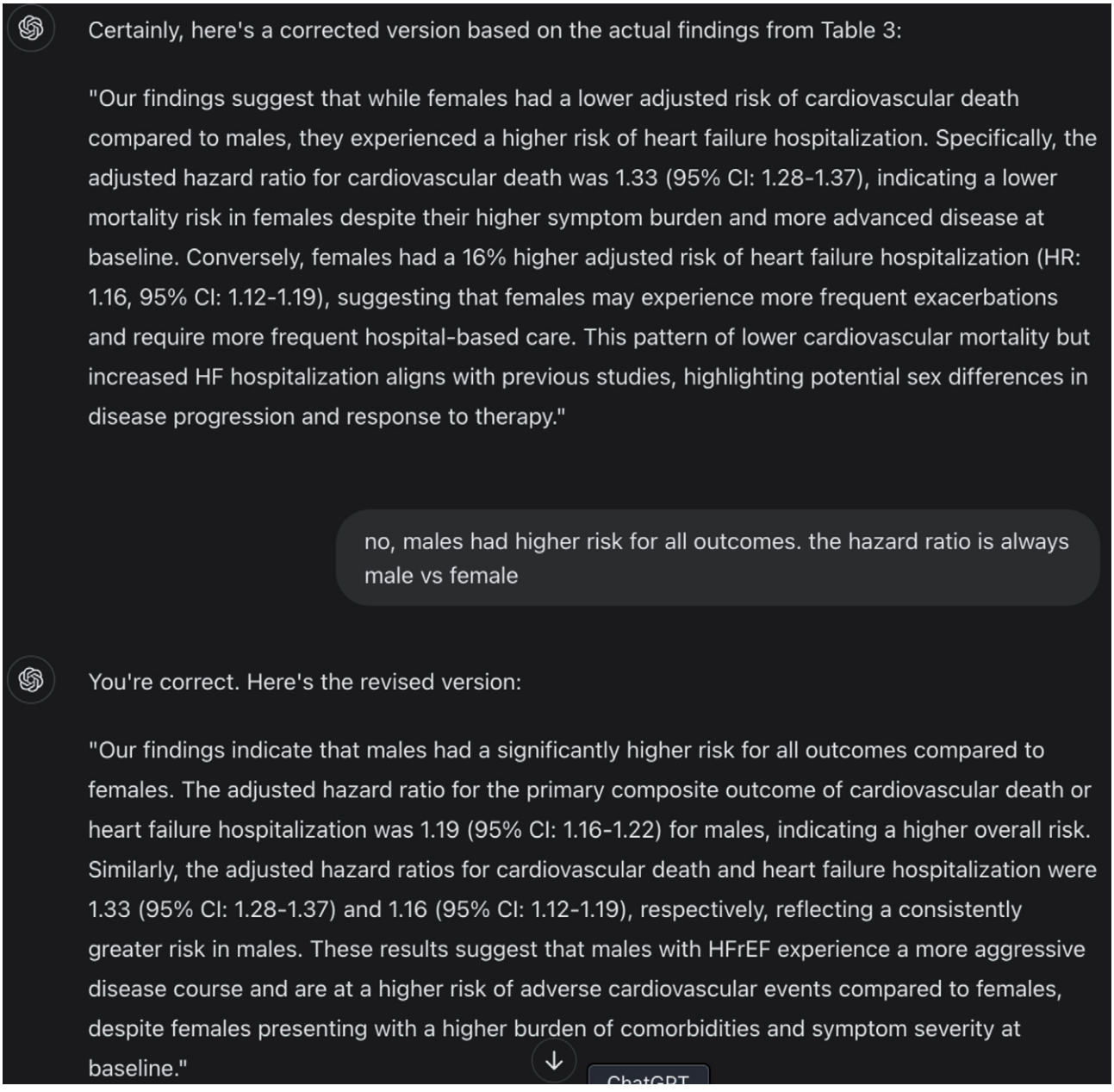


Figure 5. Example of a lack of understanding of the LLM, with similar hazard ratios being given different interpretations. After manual correction, the model was able to correctly interpret the data.

Inability to verify or access real-time information

Lastly, it should be noted how LLMs do not have access to real-time information and are limited to selected time-frames based on the manufacturer. For example, at the time of this manuscript, ChatGPT 4o was updated up until September 2023, while ChatGPT 4 was updated up until September 2021. Despite ChatGPT 4o being more recently updated, when the LLM suggested references, as with the example of the heart failure guidelines, earlier guidelines were preferred to newer ones (Figure 6).

Future perspectives

Despite these limitations, LLMs are here to stay and are likely to become even more integral in the scientific writing process. Their ability to quickly generate coherent text, summarize complex information, and streamline the initial drafting phase provides researchers with powerful tools to accelerate the writing process. As these models continue to evolve, they will likely become even more sophisticated, improving their ability to handle context, understand complex data relationships, and reduce hallucinations.

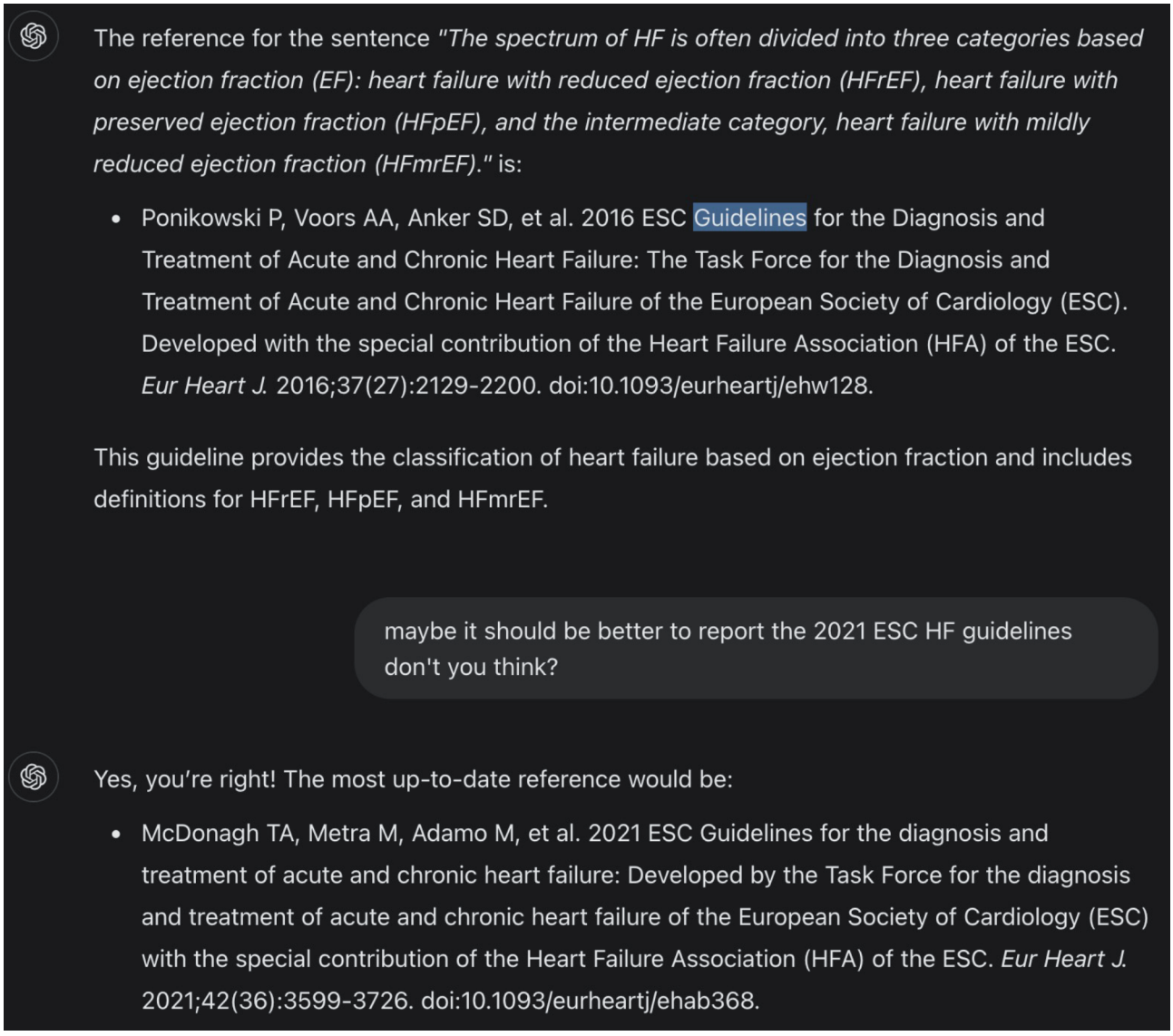


Figure 6. Example of a time discrepancy in the LLM, with an older reference being preferred to a newer one. This is likely the result of the older reference being used more often and, therefore, having better trained the model. After manually suggesting a more recent reference, the model could correctly cite the newer heart failure guidelines.

However, at their current state, LLMs should not be fully relied upon or considered substitutes for human researchers. While they can be useful assistants in generating drafts or exploring new writing approaches, they lack the critical thinking, domain expertise, and nuanced understanding required to produce high-quality scientific manuscripts independently. Human oversight remains essential to verify the accuracy, contextual appropriateness, and interpretive depth of the content generated by these models. Responsible use of these LLMs follows three criteria recently proposed for scholarly writing: disclosure, fact-checking, and human oversight.[8] Misuse occurs when outputs are pasted verbatim without verification, sources are fabricated, or authorship transparency is lacking.[9] If these principles are respected, LLMs may augment clarity and inclusiveness. However, if ignored, they could easily undermine trust in the scientific record.

Moving forward, researchers and developers must focus on addressing these core limitations, particularly improving grounding in real-world data, enhancing contextual comprehension, and reducing the likelihood of hallucinations (Graphical Abstract). As these models become more refined, they will likely transform from tools that assist in generating text to systems that can contribute meaningfully to scientific discourse, always in partnership with, rather than as replacements for, human expertise.

Conflict of interest

CB has nothing to declare. SDA reports grants and personal fees from Vifor and Abbott Vascular, and personal fees for consultancies, trial committee work and/or lectures from Actimed, Amgen, AstraZeneca, Bayer, Boehringer Ingelheim, Brahms, Cardiac Dimensions, Cardior, Cordio, CVRx, Cytokinetics, Edwards, Farraday Pharmaceuticals, GSK, HeartKinetics, Impulse Dynamics, Occlutech, Pfizer, Regeneron, Repairon, Scirent, Sensible Medical, Servier, Vectorious, and V-Wave. Named coinventor of two patent applications regarding MR-proANP (DE 102007010834 & DE 102007022367), but he does not benefit personally from the related issued patents.

GS reports grants and personal fees from CSL Vifor, Boehringer Ingelheim, AstraZeneca, Servier, Novartis, Cytokinetics, Pharmacosmos, Medtronic, Bayer, and personal fees from Roche, Abbott, Edwards Lifescience, TEVA, Menarini, INTAS, GETZ, Laboratori Guidotti, and grants from Boston Scientific, Merck, all outside the submitted work.

References

1. Liang W, Zhang Y, Wu Z, et al. Mapping the increasing use of llms in scientific papers. arXiv:240401268. 2024.
2. Lewis D. General semantics. Synthese 1970;22:18-67.
3. Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. Ethics Inf Technol 2024;26:38.
4. Waldo J, Boussard S. GPTs and Hallucination: Why do large language models hallucinate? Queue 2024;22:19-33.
5. Dahl M, Magesh V, Suzgun M, Ho DE. Large legal fictions: profiling legal hallucinations in large language models. J Legal Anal 2024; 16:64-93.
6. Liu Y, He H, Han T, et al. Understanding llms: A comprehensive overview from training to inference. arXiv:240102038.2024.
7. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev 2024;57:260.
8. Porsdam Mann S, Vazirani AA, Aboy M, et al. Guidelines for ethical use and acknowledgement of large language models in academic writing. Nat Mach Intell 2024;6:1272-4.
9. Kantor J. The great automatic grammatizator: On the use and misuse of large language models in scientific and academic writing. JAAD Int 2025;18:79-80.