Neurospine 2024;21(1):159-161.
https://doi.org/10.14245/ns.2448248.124

Neurospine
pISSN 2586-6583 eISSN 2586-6591

Commentary

Commentary on "Use of ChatGPT for Determining Clinical and Surgical Treatment of Lumbar Disc Herniation With Radiculopathy: A North American Spine Society Guideline Comparison"

Muhammad M. Abd-El-Barr

Andreas Seas[1,2], Muhammad M. Abd-El-Barr[1]

[1]Department of Neurosurgery, Duke University School of Medicine, Durham, NC, USA
[2]Department of Biomedical Engineering, Duke Pratt School of Engineering, Durham, NC, USA

Corresponding Author
Andreas Seas
https://orcid.org/0000-0003-0624-1254

Department of Biomedical Engineering, Duke Pratt School of Engineering, Hudson Hall Rm. 136, Science Dr, Durham, NC 27710, USA
Email: andreas.seas@duke.edu

See the article "Use of ChatGPT for Determining Clinical and Surgical Treatment of Lumbar Disc Herniation With Radiculopathy: A North American Spine Society Guideline Comparison" via https://doi.org/10.14245/ns.2347052.526.

Clinical medicine is a constantly changing field. However, no change is perhaps as drastic as the integration of machine learning (ML) and artificial intelligence (AI) into clinical practice. This rapid adaptation has recently been stretched with the introduction of the chat generative pre-trained transformer (ChatGPT) in 2022. Unlike many other complex tools for ML, ChatGPT is a large language model (LLM) developed with the intent for rapid use by the lay audience. The tremendously low barrier to entry—namely involving generation of an account—has led to expansive interest in the use of ChatGPT in nearly every subfield of surgery, including spine surgery and low back pain. The goal of the study by Mejia et al.[1] was to assess the ability of ChatGPT to provide accurate medical information regarding the care of patients with lumbar disk herniation with radiculopathy.

The research team developed a series of questions related to lumbar disk herniation, using the 2012 North American Spine Society (NASS) guidelines as a gold standard.[2] They then collected responses from both ChatGPT-3.5, and ChatGPT-4.0. They quantified several metrics for each response. A response was considered accurate if it did not contradict the NASS guidelines. It was considered overconclusive if it provided a recommendation when the NASS guidelines did not provide sufficient evidence. A response was supplementary if it included additional relevant information for the question. Finally, a response was considered incomplete if it was accurate but omitted relevant information included within the NASS guidelines.

Both ChatGPT-3.5 and -4.0 provided accurate responses to just over 50% of questions. Nearly half of all responses were also overconclusive, providing recommendations without direct backing of the NASS guidelines. Interestingly, both models provided supplemental information in most of their responses yet were also noted to have provided incomplete responses to 11/29 and 8/29 questions for ChatGPT-3.5 and -4.0, respectively.

At face value, these findings indicate that both ChatGPT models provided inaccurate and overconclusive recommendations in the context of lumbar disk herniation with radiculopa-

thy. However, the recommendations from NASS 2012 did not account for evidence from the following decade of research which may have been considered in the responses generated by ChatGPT. To assess this, the authors looked at several of the recommendations generated by ChatGPT which were either inconsistent with the NASS 2012 guidelines or classified as overconclusive. In doing so, they found that ChatGPT appeared to have extrapolated several heuristics from more recent literature. These included (1) lower risk of infection at ambulatory surgery centers, (2) reduced costs of microdiscectomy in the ambulatory setting, and (3) reduced complication rates from full endoscopic lumbar discectomy as compared to open discectomy/microdiscectomy. While there is some evidence for each of these heuristics, they all represent generalizations of extremely complex systems. While the authors do mention that ChatGPT "duly recognized" the limits of these heuristics, it is unclear how this was conveyed in the final response, and whether a lay reader could have understood these caveats.

The salient message from these data is that both ChatGPT models cannot reliably provide accurate recommendations for the management of lumbar disk herniation with radiculopathy. Furthermore, both often provide overconclusive recommendations which appear to be extrapolated from literature published after 2012. This tendency reflects a potentially dangerous phenomenon among LLMs: the ability for them to "hallucinate." A LLM "hallucination" takes place when the model's response to a question includes inaccurate conclusions or assertions. It can be the result of (1) inaccurate or contradictory source material, (2) missing data, (3) the model's variability parameter (often called its "temperature"), or any combination of the above.

There are several ways to mitigate LLM hallucination, which are applicable in the future use of ChatGPT as a potential clinical tool. The first is the use of reinforcement learning from human feedback, a paradigm wherein models utilize feedback from users in real-time to fine-tune their text-generation parameters.[3] Another important method is retrieval augmented generation (RAG), a technique wherein a "retriever" pulls data from a relevant corpus of knowledge to optimize the prompt fed to the generative engine behind GPT or any other LLM.[4] The RAG architecture has recently seen application in neurosurgery with the creation of AtlasGPT.[5] Yet another approach involves the use of data source "weights" to assign a degree of trust to each source: some sources such as peer-reviewed scientific literature could carry greater weight than text from pharmaceutical advertising websites. The challenge herein resides in the vast volume of data to be weighted, a task which may need its own complex LLM. Finally, model "temperature," the parameter associated with the variability can be adjusted to minimize hallucination.

This study clearly outlines the limits of using a general LLM like ChatGPT to help guide patient care without any adjustments. However, there are several ways this work could have been improved to provide further insight into the development of future tools for guiding patient care in the spine clinic and ward. Firstly, the authors utilized prompts that matched nearly word-for-word with NASS 2012 guidelines. This allowed them to assess the model's ability to regurgitate guidelines, but failed to demonstrate how ChatGPT would respond to realistic clinical questions from patients and physicians. Furthermore, they did not attempt to perform prompt engineering, the practice of optimizing the way an LLM is queried to generate clear results.[6] Without rigorous prompt engineering, even the best LLMs can provide ambiguous, or biased results, rely too heavily on patterns within training data, or even entirely misinterpret the intent of the user's question. The authors note this when asking ChatGPT on the "value of treatment," and the model assumed the reader was asking about the relative value of different surgical procedures. Rather than using prose questions taken from the NASS guidelines, future work could utilize descriptions of patient or physician queries organized within a custom prompt optimized through several rounds of prompt engineering using common patterns described in the LLM literature.[6]

Another limitation of this work was the use of the ChatGPT online interface, rather than its application program interface. While the use of the interface does better reflect the most common interface used by physicians and patients, it also prevented the authors from testing model output stochasticity by varying its "temperature." A final limitation herein was the fact that the NASS 2012 guidelines may have been used as elements of the ChatGPT-3.5 and -4.0 training sets. This could similarly be prevented with the use of user-generated prose addressing NASS guidelines, without the use of similar or identical questions and text.

The world of clinical medicine has entered a new renaissance with the advent of ML tools like ChatGPT. This work demonstrates that rapid growth in the clinical application of AI comes with significant risks, especially when tools like ChatGPT are so readily accessible by patients and physicians. It is crucial that all healthcare workers, whether they are actively engaged in AI work or not, to use care in their use of LLMs and their conversations with patients on this new technology.[7] As this technology becomes more mature, it will be interesting to see if these

 https://doi.org/10.14245/ns.2448248.124

models will start to 'outperform' our benchmarks of clinical care guidelines, controlled studies and 'clinical judgement.'

• **Conflict of Interest:** The authors have nothing to disclose.

REFERENCES

1. Mejia MR, Arroyave JS, Saturno M, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine 2024;21:149-58.
2. Kreiner DS, Hwang SW, Easa JE, et al. An evidence-based clinical guideline for the diagnosis and treatment of lumbar disc herniation with radiculopathy. Spine J 2014;14:180-91.
3. Stiennon N, Ouyang L, Wu J, et al. Learning to summarize from human feedback. arXiv:2009.01325v3 [Preprint]. 2022 [cited 2024 Mar 4]. Available from: https://doi.org/10.48550/arXiv.2009.01325.
4. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv:2005.11401v4 [Preprint]. 2021 [cited 2024 Mar 4]. Available from: https://doi.org/10.48550/arXiv.2005.11401.
5. Hopkins BS, Carter B, Lord J, et al. Editorial. AtlasGPT: dawn of a new era in neurosurgery for intelligent care augmentation, operative planning, and performance. J Neurosurg 2024 Feb 27:1-4. doi: 10.3171/2024.2.JNS232997. [Epub].
6. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382v1 [Preprint]. 2023 [cited 2024 Mar 4]. Available from: https://doi.org/10.48550/arXiv.2302.11382.
7. Dorr DA, Adams L, Embí P. Harnessing the promise of artificial intelligence responsibly. JAMA 2023;329:1347-8.