ASRA PAIN MEDICINE

49th Annual Regional Anesthesiology and Acute Pain Medicine Meeting
March 21-23, 2024 | San Diego, California | #ASRASPRING24

Abstract: 5267

Scientific Abstracts > Emerging Technology

# Large Language Models in Medical Literature Review: Automating the Analysis of Erector Spinae Plane Block Studies

Hyo Jung Hong, Lu Yang, Sharon Chao, Ban Tsui
Stanford University

## Introduction

The accelerating growth of biomedical literature presents a challenge for clinicians to stay current with evidence, especially in regional anesthesia where the annual publication number has grown exponentially.1 According to the National Library of Medicine, nearly 1.4M new citations were added to MEDLINE in 2022 alone. This is particularly evident in the field of regional anesthesiology, where literature around newer blocks are rapidly evolving. Systematic literature reviews are becoming increasingly cumbersome to manually conduct due to the growing volume of new publications. Large language models (LLMs), such as ChatGPT developed by OpenAI, are demonstrating improving performance in various natural language processing (NLP) related tasks and scientific knowledge.2 Generative artificial intelligence (AI) models can now perform question-and-answering (QA) based on a given context.3 While multiple studies have evaluated LLMs on standardized medical examinations, LLM performance on evidence-based knowledge on a clinical subspecialty has been understudied.4,5 This study aims to assess the readiness of ChatGPT to conduct an automated systematic review of the literature on erector spinae plane block (ESPB), and determine if its responses are comparable to those obtained through a manual literature review.

## Materials and Methods

### Manual Review

Prior to comparing the outputs of the ChatGPT models, a set of standard answers was established and generated through a manual systematic literature review of ESPB-related publications from August 1, 2018 to December 31, 2022. Questions that were explored are summarized in Table 1. The manual study is detailed in an abstract separately submitted to ASRA (#5371) with the title, “The Erector Spinae Plane Block: A pooled data of 6,495 cases since its first publication” (Chao et al.).

### Automated Review Using LLM

For the model, we utilized ChatGPT, supported by OpenAI's text-embedding-ada-002 model, for an automated review that mirrors our manual review process. This involved a streamlined three-step approach: First, we processed and extracted texts from 608 publications that met our inclusion criteria, originally in PDF format. Then, we transformed these texts into numerical vector embeddings, creating a knowledge base for the LLM. Finally, we inputted the manual review questions for QA against the constructed vector database and obtained the model’s responses, enabling a direct comparison with the gold standard answers. For clarification on technical terms used, refer to Table 2, and the complete methodology is illustrated in Figure 1.

### Model Evaluation

The LLM results were evaluated in comparison to manual review responses, which functioned as our reference point for factual information. To evaluate the alignment between the LLM-generated responses and the reference (manual review answers), we employed the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric. This method quantifies the overlap of “n-grams” between the two sets of responses. Furthermore, we calculated F1 scores, a metric that integrates both precision and recall, to gauge the LLM's performance in relation to the gold standard answers.

## Results/Case Report

Our preliminary results indicated that ChatGPT’s performance in answering questions, when compared to the gold standard answers from the manual systematic literature review, was generally suboptimal. F1 scores as detailed in Table 1, varied significantly across different questions, ranging from 0.00 to 0.58. In our qualitative examination of the model’s outputs, exemplified in Figure 1, we observed three patterns in its performance. Firstly, for certain questions, the model either expressed an inability to provide an answer or indicated a lack of sufficient context (Figure 2, question 2). Secondly, the model showed moderate accuracy responding to questions about the percentage of patients receiving multimodal analgesia, the comparison between single shot blocks and catheter-based methods, and the most common anatomic locations for ESPB (Figure 1). However, for several questions, the model’s answers deviated significantly from the gold standard answers (Figure 2, question 5), with errors exceeding 280%.

## Discussion

This research highlights the constraints of ChatGPT in its current iteration or format when it comes to employing LLMs for specialized medical knowledge domains, such as Erector Spinae Plane Block (ESPB), despite considerable interest in the application of AI in the medical field, particularly for tasks like medical literature search. Presently, it is evident that a significant disadvantage associated with the use of LLMs in the medical field is that these models are susceptible to hallucinations, which involves the generation of false outputs that are statistically plausible. To tackle this issue, our research endeavored to contextualize peer-reviewed ESPB literature within a popular LLM using OpenAI's embedding retrieval. Despite this approach, the low F1 scores indicate ChatGPT’s overall poor performance in answering questions, even with access to expert-selected literature as a knowledge base. However, translating F1 scores to model reliability is challenging, given ChatGPT’s linguistic variability . For instance, despite some accurate responses in questions like Q3 and Q4 (Figure. 2), their F1 scores remained low, between 0.16 and 0.47.

While the performance of ChatGPT falls short of expectations, it should be emphasized that the uniqueness of this investigation stems from its methodology. First, the papers and questions used in the LLM analysis were manually selected by clinical experts. Second, the study specifically targets a particular medical field in order to assess the LLM in systematic literature review. Additionally, it benchmarks the LLM's results with answers obtained from manual systematic literature reviews conducted by clinical experts.

However, this study is not without limitations. It relies on the accuracy of manual systematic reviews as a benchmark, potentially overlooking their subjective nature. Furthermore, the challenges in interpreting F1 score discussed above underline the necessity for a more holistic set of metrics to assess LLMs in clinical medicine, where information exchange is complex, contextual, and often incomplete. Future evaluations should consider additional metrics like accuracy, calibration, robustness, bias, and efficiency to ascertain clinical applicability. The resources and costs involved in developing, evaluating, and updating these models also pose practical challenges. Future research should compare open-source models and the models trained on both open-access and restricted-access papers, to evaluate the impact of data diversity on model performance, enhancing their utility in clinical practice.

In conclusion, the study highlights the critical role of physicians in guiding and enhancing the use of LLM in healthcare settings.

## References

1. Shbeer A: Regional Anesthesia (2012–2021): A Comprehensive Examination Based on Bibliometric Analyses of Hotpots, Knowledge Structure and Intellectual Dynamics. J Pain Res 2022; 15:2337–50
2. Mao R, Chen G, Zhang X, Guerin F, Cambria E: GPTEval: A Survey on Assessments of ChatGPT and GPT-4 2023 at <http://arxiv.org/abs/2308.12488>
3. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, Sun L: A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT 2023 doi:10.48550/ARXIV.2303.04226
4. Kung TH, Cheatham M, Medenilla A, Sillos C, Leon LD, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2:e0000198
5. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, Asaad WF, Cielo D, Oyelese AA, Doberstein CE, Gokaslan ZL, Telfeian AE: Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery 2023; 93:1353–65

## Disclosures

No

## Tables / Images

□

□

□

□