

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

Hyo Jung Hong, MD, MBA, Sharon Chao, BA, Lu Yang, PhD, Ban Tsui, MD
Department of Anesthesiology, Perioperative, and Pain Medicine
Stanford University School of Medicine, Stanford, CA



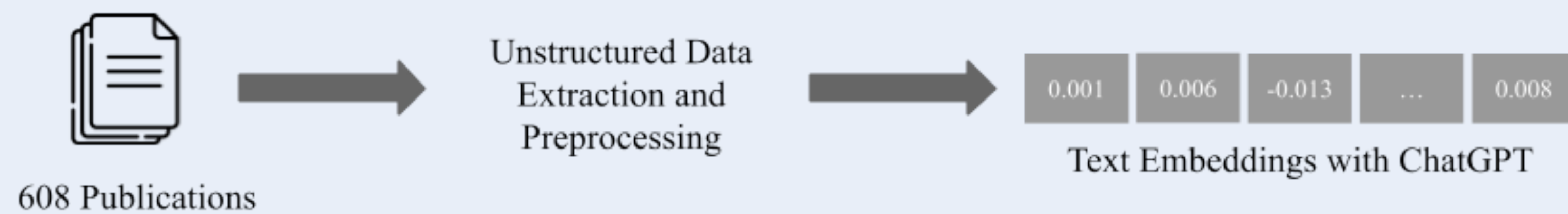
Background

- **Literature is growing fast**, with over 1.4M citations added to MEDLINE in 2022 alone¹
- **Large language models (LLMs)**, such as ChatGPT, are demonstrating improving performance in processing vast amounts of texts and question-and-answering.^{2,3}
- **How well can ChatGPT conduct automated systematic review of Erector Spinae Plane block (ESPB) literature relative to traditional manual reviews?**

Methods

1. **Manual review** of ESPB literature published between August 1, 2018 and December 31, 2022 (gold standard answers)
2. **Automated review** of the same literature was done via advanced text-embedding techniques in ChatGPT
3. **Comparison** of ChatGPT outputs to the gold standard answers

Preparing the knowledge base



Automating the literature review

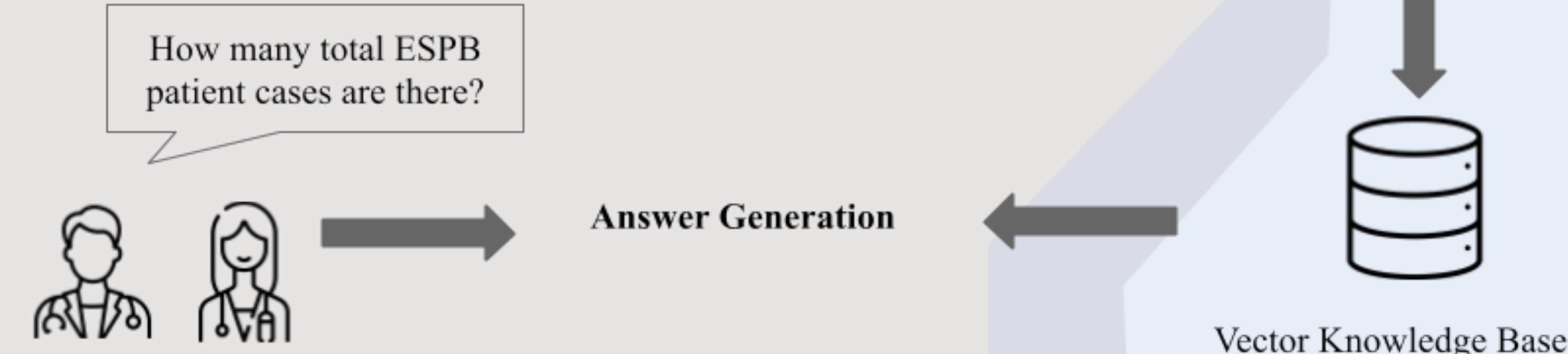


Figure 1: Overview of methodology for automated literature review using LLM

Results

Table 1: Accuracy and precision metrics of ChatGPT outputs relative to gold standard

Select Questions, Gold Standard (GS) Answers, and Model Outputs	ROUGE-1	Precision	F1
Q. Are there studies done comparing ESPB to another block? Yes or No. GS: Yes. Model: Yes. Three studies compared ESPB to other blocks such as the oblique subcostal transverse abdominis plane block and PVB.	0.500	0.053	0.095
Q. What percentage of patients studied received multimodal analgesia? GS: 90.9%. Note that 24.2% of cases did not report whether multimodal analgesia was used, while only 2.5% reported no multimodal analgesia. Model: In this review, 90.9% of patients reported the use of multimodal analgesia in addition to the ESPB.	0.261	0.375	0.308
Q. In percentage, what were the most common three anatomic locations of ESPB? GS: The most common three anatomic ESPB locations were thoracic 86.3%, followed by lumbar 13.1%, and cervical 0.3%. Model: According to the provided information, the most common three anatomic locations of ESPB were thoracic (89.9%), cervical (0.8%), and unknown (7.9%).	0.500	0.435	0.465
Q. For single shot ESPB, what were the most commonly used local anesthetic (LA) name, percentage, and volume administered? GS: Bupivacaine 0.25% was the most commonly used LA with a mode of 20 mL. Ropivacaine 0.5% was the next most commonly used LA with a mode of 20 mL. Model: The most commonly used local anesthetic (LA) names for single shot ESPB were lidocaine and mepivacaine. The percentage of cases in which these LA were used as additives was 10.7%. The volume administered ranged from 10 mL to 30 mL, with a mode range of 20 mL-30 mL.	0.579	0.262	0.361
Q. What were the most commonly used adjuvant medications added to the local anesthetics in single shot ESPB? Provide the percentage of each additive use across all reported single shot ESPB. GS: Steroids (10.3%), dexmedetomidine (9.3%), epinephrine (4.5%), and local anesthetics (2.2%) were the most commonly used additives in single shot ESPB cases Model: The most commonly used adjuvant medications added to the local anesthetics in single shot ESPB were steroids, local anesthetics (lidocaine, mepivacaine), and epinephrine. These additives were used in 10.7% of reported single shot ESPB cases.	0.625	0.536	0.577

- ChatGPT generally **underperformed** in conducting systematic literature review of ESPB literature, exhibiting moderate accuracy to notable discrepancies from gold standard answers.
- It performed well on answering factual questions at the article level and poorly on questions requiring aggregating studies.

Discussion

- The study highlights the **limitations of current ChatGPT** in performing literature review on a clinical subspecialty topic.
- The study uniquely uses a **manual review as benchmark** for the LLM performance and focuses on a **subspecialty topic (ESPB)**.
- Future directions include developing **more holistic LLM evaluation metrics** tailored to clinical medicine and evaluating different LLMs for biomedical literature review.
- We emphasize the **important role of physicians** in applying and validating LLM capabilities in clinical use cases.

Download the Abstract!



← Scan with phone camera to download the entire abstract

References

- 1.Shbeer A: Regional Anesthesia (2012–2021): A Comprehensive Examination Based on Bibliometric Analyses of Hotspots, 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