

Improving Performance of Large Language Models on Anesthesiology Topics



BALAJI PANDIAN, MD, MBA¹; ROHAN JOTWANI, MD, MBA¹; JOHN RUBIN, MD¹

¹Department of Anesthesiology, New York Presbyterian – Weill Cornell, New York, NY

INTRODUCTION

- Large Language Models (LLMs) have gained widespread popularity
- ChatGPT** (OpenAI Inc., San Francisco, US) was released in November 2022
 - 100 million active users** within 2 months
 - 1,500+ PubMed indexed articles** within the first year
 - Passing benchmark for the United States Medical Licensing Exam (USMLE) [1]
- Performance in Anesthesiology is similar
 - 56.2% on American Board of Anesthesiology (ABA) Board Examination preparation book**
 - Noted variability between subjects: 30.3% on “anesthetic equipment” question and 64.8% on “knowledge recall” questions [2]
- “Fine-tuning” existing LLMs and utilizing “document-augmented retrieval” can improve performance on domain-specific knowledge
 - Chen et al. have demonstrated improvements of 10-15% on USMLE-style questions after fine-tuning LLMs on exam questions, PubMed Abstracts, and clinical guidelines [3]
- In this study, we demonstrate a marked improvement in accuracy on a previous regional anesthesia question set from Anesthesia Toolbox [4] using unblinded document-augmented retrieval on articles from OpenAnesthesia

METHODS

- ChatGPT from OpenAI was programmatically accessed (12/15/2023 - Python 3.9.13 on Mac OS 12.5) to insert and evaluate input prompts
- “Select the single best answer:” was prepended to each of the 36 regional anesthesia question stems to obtain a succinct and unambiguous response
- Incorrect answers from the original ChatGPT evaluation were categorized through human anesthesiologist sorting into a 3-6 word subject that encompassed the high-level learning objective of the question
- OpenAnesthesia Keyword Articles [5] (if available) for each of the subjects were translated from HTML to text and fed into the ChatGPT “Knowledge” database before being re-evaluated with this new model

RESULTS

Anesthesia Toolbox

“Regional Anesthesia” Questions

```
{
  "PREFIX_PHRASE": "Select the single best answer:",
  "QUESTION_STEM": "Which of the following nerve blocks is most appropriate for a total knee replacement?",
  "ANSWER_OPTIONS": {
    "A": "Adductor Canal",
    "B": "Sciatic",
    "C": "Deep Peroneal",
    "D": "Transversus Abdominis Plane"
  }
}
```

Original ChatGPT



Correct	Incorrect
n=19/36	n=17/36



Article Found	No Article
n=9/17	n=8/17

ChatGPT
 OpenAI

+



Total Knee Arthroplasty: Regional Anesthesia Techniques

Last updated: 01/12/2024

Key Points

- Total knee arthroplasty (TKA) is a commonly performed procedure, with its demand expected to increase to over 1 million procedures by the year 2030. Over the last decade, TKAs have shifted toward outpatient surgery. Postoperative pain control among TKA patients continues to be an area for improvement.
- In the last decade, there has been a steady shift toward outpatient surgery. In 2018, the United States Center for Medicare & Medicaid Services finalized the 2018 Medicare Outpatient Payment System rule that removed TKA from the Medicare inpatient-only list of procedures.
- In the last decade, there has been a steady shift toward outpatient surgery. In 2018, the United States Center for Medicare & Medicaid Services finalized the 2018 Medicare Outpatient Payment System rule that removed TKA from the Medicare inpatient-only list of procedures.
- In addition to optimizing pain control, the other important consideration is providing analgesia while preserving motor function to facilitate immediate postoperative participation in physical therapy.

Background

- TKA is commonly performed to treat end-stage osteoarthritis and other disabling knee conditions, and its demand is expected to increase to over 1 million procedures in the United States by the year 2030.
- In the last decade, there has been a steady shift toward outpatient surgery. In 2018, the United States Center for Medicare & Medicaid Services finalized the 2018 Medicare Outpatient Payment

Example Article (1/9)

	Correct	Incorrect
Original ChatGPT	52.7% (n=19/36)	47.3% (n=17/36)
ChatGPT + OpenAnesthesia	72.2% (n=26/36)	27.8% (n=10/36)

ChatGPT with OpenAnesthesia Article knowledge achieved an overall accuracy of 72.2%, a 19.5% improvement from the original model

Figure 1. Data and Model Assessment: OpenAnesthesia topics queried and used as knowledge for incorrect questions. ChatGPT with unblinded background knowledge performs markedly better.

DISCUSSION

- ChatGPT struggles with niche topics such as regional anesthesia and acute pain**
- Creating a domain-specific LLM from scratch would require not only substantial computational effort, but also require a corpus of validated input data and text which does not currently exist**
 - Additionally, Beyond board preparations questions, no accepted benchmark exists for model evaluation
- Document-augmented retrieval and fine-tuning LLMs may offer an effective methodology to optimize LLM performance on niche topics**
- Notable limitation of our study is the unblinded nature of document-retrieval on inaccurate questions only**
- Next steps to scrape the entire compendium of OpenAnesthesia Keyword Articles (approx. 1000)**

REFERENCES

- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230.
- Shay D, Kumar B, Bellamy D, Palepu A, Dershwitz M, et al. Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. Br J Anaesth. 2023 Aug;131(2):e31-e34. doi: 10.1016/j.bja.2023.04.017. Epub 2023 May 18. PMID: 37210278.
- Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, et al. Meditron-70B: Scaling Medical Pretraining for Large Language Models. arXiv:2311.16079v1. November 27, 2023.
- Pandian B, Burns M, Rubin J, Pryor K, Brumberger E, Jotwani R. ChatGPT Performance Versus Residents on Multiple-Choice Regional and Chronic Pain Assessments. American Society of Regional Anesthesia and Pain Medicine (ASRA). Spring ASRA 2023 Meeting. Hollywood, FL. Presented April 22, 2023.
- “OpenAnesthesia: Keywords.” International Anesthesia Research Society. Accessed December 15, 2023. <https://www.openanesthesia.org/keywords/>

Question/Comments? Please contact: Balaji Pandian at bap9037@nyp.org