Comparative Analysis of Popular AI-Chat Bots in Generating Inguinal Hernia Surgery Recommendations

Sumaya Bokhari[1]; Ashley Tran, MD[1]; Thisun Udagedara[1]; Sharon Shiraga, MD[1]; Stuart Abel, MD[1]; Caitlin Houghton, MD[1]; Katie Galvin, MD[1]; Kamran Samakar, MD[1]; Luke R Putnam, MD[1]

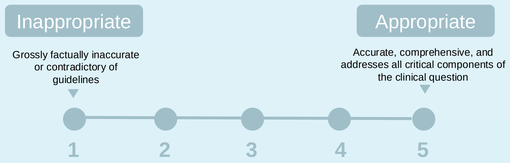
1. Division of Upper GI and General Surgery, Department of Surgery, Keck Medical Center of USC

Introduction

- Artificial intelligence (AI), particularly large language models (LLMs), has garnered increasing interest for its potential clinical applications.
- LLMs have been shown to produce appropriate responses to questions in various medical fields.
- However, minimal data exists evaluating the use of LLMs within the field of inguinal hernia repair (IHR).
- The aim of this study was to evaluate and compare the accuracy and comprehensiveness of responses generated by 3 popular LLMs to evidence-based clinical questions regarding the management of inguinal hernias.

Methods

- Questions were generated from recommendations with strong levels of evidence included in the international HerniaSurge guidelines for groin hernia management.
- Questions were categorized into 4 clinical areas – Diagnosis, Perioperative Care, Surgical Management, and Other.
- LLM-generated responses to questions were recorded and blinded for review by 6 general surgeons.
- Response quality was graded using a 5-point Likert scale:



- Readability of LLM responses was evaluated using 6 validated readability formulae.

Results

Mean Response Grading

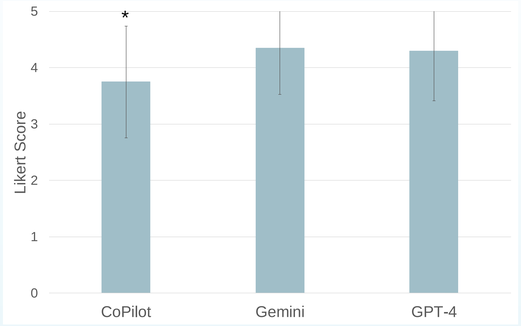


Figure 1. Overall mean response grading

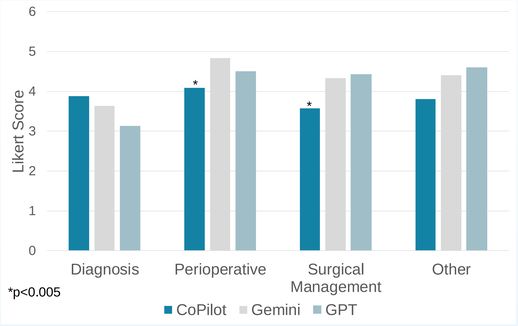


Figure 2. Mean response grading by category

- A total of 30 questions were compiled, resulting in 90 LLM responses across the 3 models.
- On average, responses generated by CoPilot were graded significantly lower than responses generated by Gemini or Chat GPT4 (Figure 1) (CoPilot: 3.75±0.99, Gemini: 4.35±0.82, GPT4: 4.30±0.89, p<0.001).
- On subanalysis by question category, Gemini and GPT4 outperformed CoPilot in generating responses to questions about Perioperative Care and Surgical Management (Figure 2).

Readability

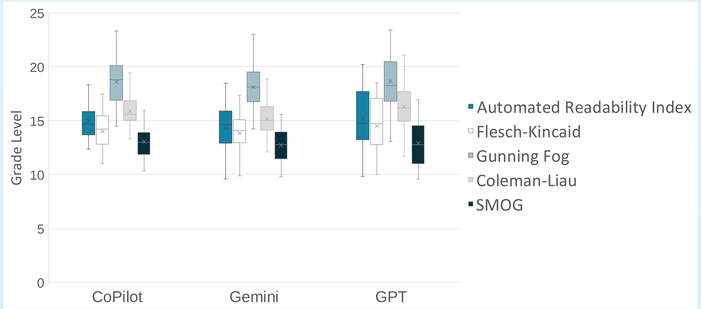


Figure 3. Readability scores

- There was no significant difference in readability between the responses generated by the different LLMs
- All 3 LLMs produced responses with high reading levels, ranging from college to college-graduate reading levels
- Figure 3 summarizes the mean readability scores for LLM responses using validated readability formulas

Discussion

- This study demonstrated variability in the performance of different AI-powered LLMs in generating appropriate responses to questions about IHR.
  - Gemini and Chat GPT-4 outperformed CoPilot overall, particularly in questions regarding perioperative care and surgical management.
- Readability was consistently difficult across all 3 models
  - Reading levels ranged from college to college graduate level
  - Poses a challenge to accessibility given average reading level in the US is 6th to 8th grade

Implications

- AI tools such as LLMs have the potential to transform healthcare workflows and provide patients with adjunctive tools for obtaining health information.
- However, further research is needed to address challenges to adoption such as improving accessibility and optimizing accuracy.

References

1.Lee Y, Shin T, Tessier L, et al. Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations. *Surg Obes Relat Dis*. 2024;20(7):603-608. doi:10.1016/j.soard.2024.03.011

2.Leng Y, Yang Y, Liu J, Jiang J, Zhou C. Evaluating the Feasibility of ChatGPT-4 as a Knowledge Resource in Bariatric Surgery: A Preliminary Assessment. *Obes Surg*. 2025;35(2):645-650. doi:10.1007/s11695-024-07666-8

3.Rooney MK, Santiago G, Perni S, et al. Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *J Patient Exp*. 2021;8:2374373521998847. Published 2021 Mar 3. doi:10.1177/2374373521998847