Surgical Endoscopy
https://doi.org/10.1007/s00464-024-11155-5

Clinical artificial intelligence: teaching a large language model to generate recommendations that align with guidelines for the surgical management of GERD

Bright Huo[1] · Nana Marfo[2] · Patricia Sylla[3] · Elisa Calabrese[4] · Sunjay Kumar[5] · Bethany J. Slater[6] · Danielle S. Walsh[7] · Wesley Vosburg[8]

Received: 24 May 2024 / Accepted: 4 August 2024

Abstract

Background Large Language Models (LLMs) provide clinical guidance with inconsistent accuracy due to limitations with their training dataset. LLMs are "teachable" through customization. We compared the ability of the generic ChatGPT-4 model and a customized version of ChatGPT-4 to provide recommendations for the surgical management of gastroesophageal reflux disease (GERD) to both surgeons and patients.

Methods Sixty patient cases were developed using eligibility criteria from the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) & United European Gastroenterology (UEG)-European Association of Endoscopic. Surgery (EAES) guidelines for the surgical management of GERD. Standardized prompts were engineered for physicians as the end-user, with separate layperson prompts for patients. A customized GPT was developed to generate recommendations based on guidelines, called the GERD Tool for Surgery (GTS). Both the GTS and generic ChatGPT-4 were queried July 21st, 2024. Model performance was evaluated by comparing responses to SAGES & UEG-EAES guideline recommendations. Outcome data was presented using descriptive statistics including counts and percentages.

Results The GTS provided accurate recommendations for the surgical management of GERD for 60/60 (100.0%) surgeon inquiries and 40/40 (100.0%) patient inquiries based on guideline recommendations. The Generic ChatGPT-4 model generated accurate guidance for 40/60 (66.7%) surgeon inquiries and 19/40 (47.5%) patient inquiries. The GTS produced recommendations based on the 2021 SAGES & UEG-EAES guidelines on the surgical management of GERD, while the generic ChatGPT-4 model generated guidance without citing evidence to support its recommendations.

Conclusion ChatGPT-4 can be customized to overcome limitations with its training dataset to provide recommendations for the surgical management of GERD with reliable accuracy and consistency. The training of LLM models can be used to help integrate this efficient technology into the creation of robust and accurate information for both surgeons and patients. Prospective data is needed to assess its effectiveness in a pragmatic clinical environment.

Keywords GERD · Surgery · ChatGPT · Natural language processing · Large language models · Guidelines · Artificial intelligence

✉ Wesley Vosburg
wesvosburg@gmail.com

[1] Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

[2] Ross University School of Medicine, Miramar, FL, USA

[3] Division of Colon and Rectal Surgery, Department of Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4] University of Adelaide, Adelaide, SA, Australia

[5] Department of General Surgery, Thomas Jefferson University Hospital, Philadelphia, PA, USA

[6] Department of Surgery, University of Chicago, Chicago, IL, USA

[7] Department of Surgery, University of Kentucky, Lexington, KY, USA

[8] Department of Surgery, Mount Auburn Hospital, Harvard Medical School, Cambridge, MA, USA

Published online: 12 August 2024

Since their introduction, large language models (LLMs) have gained popularity due to widespread accessibility and impressive ability to generate prompt answers using expert or layperson language [1]. Recognizing this, patients are beginning to use LLMs for clinical advice for various topics related to their health. Despite tremendous interest, the clinical application and endorsement of the use of LLMs within healthcare remains restricted [2]. While LLMs have proven valuable in various domains [3], they have exhibited inconsistency in providing medical advice and recommendations. A major barrier to their acceptance in clinical practice is the clinical accuracy of their responses [4]. LLM-generated recommendations can lack evidence-based support or offer information inconsistent with clinical guidance.

LLMs such as ChatGPT (Open AI) are trained on extensive datasets [5] on various topics, but its training data lacks specificity to medical contexts. With growing interest in the use of LLMs for clinical application, there is a need for developing tailored LLM models that are medically relevant. However, developing a LLM is time-consuming and resource-intensive, requiring unique expertise. Recently, ChatGPT-4 released a "customize GPT" function that enables the user to direct the responses of the LLM for specific purposes. The ability to tailor a LLM toward utilizing evidence-based recommendations as a primary data source to mitigate clinical inaccuracies and inconsistencies may provide an opportunity to significantly enhance their reliability.

Tailoring pre-trained LLMs which have been extensively trained on diverse datasets for medical applications offers a promising strategy [6]. These LLMs would be able to offer clinically accurate recommendations with a short induction period of "customization," avoiding the resource and time constraints associated with developing a LLM from scratch. The aim of this study was to develop a customized ChatGPT using guidelines to create a LLM-linked chatbot that provides accurate clinical recommendations and compare it to an untrained GPT model.

Materials and methods

Objective, model customization, & prompt engineering

On July 21st, 2024, ChatPT-4's "Create a GPT" feature was used to customize a version of ChatGPT for our purposes. A paucity of literature exists on the reliability of ChatGPT-4's customization feature and the number of prompts needed to ensure its reliability. First, the model was informed that its purpose would be to guide clinicians on the surgical management of gastroesophageal reflux disease (GERD) based on the SAGES & UEG-EAES clinical practice guidelines [7]. We chose these guidelines as our team members have extensively reviewed the randomized evidence behind the surgical management of GERD. We further trained the model using the 2022 UEG/EAES guideline recommendations on the surgical management of GERD [8]. The model was told to use layperson language for patients, but to use conventional medical terminology and a professional tone when conversing with surgeons. Secondly, a PDF copy of the SAGES & UEG-EAES clinical practice guideline on the surgical management of GERD was uploaded to the website [7]. Examples were given to the model regarding patient cases and clinical questions that applied to the first two key recommendations from the 2021 guidelines. Specifically, hypothetical patient cases were posed to the custom model in-progress, asking whether patients should receive surgery, or whether they should receive surgery robotically or laparoscopically. Thirdly, feedback on whether the model's response was correct was provided iteratively. For instance, the model was corrected to make a firm recommendation based on the guideline recommendations. Steps two and three comprised our prompt engineering/testing phases. This iterative input and feedback loop was repeated until the model provided a correct response aligning with the guideline recommendations for three consecutive cases for this pilot study. This process was completed by the lead author over 1.5 h. During model training, only four clinician-oriented cases and questions were posed. No patient-oriented questions were posed. Additional queries were inputted to the generic ChatGPT-4 to ensure that prompts were structured appropriately to elicit responses from both models to complete the prompt engineering/testing phase. (Fig. 1).

Query strategy

Standardized patient cases were developed based on key questions from the SAGES & UEG-EAES guidelines for the surgical treatment of GERD [7]. These cases specified combinations of patient age, clinical history, and clinical questions based on the relevant guideline recommendations. Each case reflected the population, intervention, and comparator addressed by the applicable key question. With input from practicing general and foregut surgeons, clinical question phrasing was refined. These cases were used as prompts to query our customized version of ChatGPT-4 as well as the generic version of ChatGPT-4 on July 21st, 2024 from a computer serve in Hamilton, Ontario, Canada (Table 1). The most recent update to the generic ChatGPT-4 model was May 13th, 2024.

Patient prompts were generated based on surgeon prompts by adjusting phrasing to reflect layperson terminology while limiting medical terminology. No follow-up prompts or medical disclaimers were applied. No prompts contained any reference to professional organizations, societies, or countries. All prompts were constructed in English. All prompts were

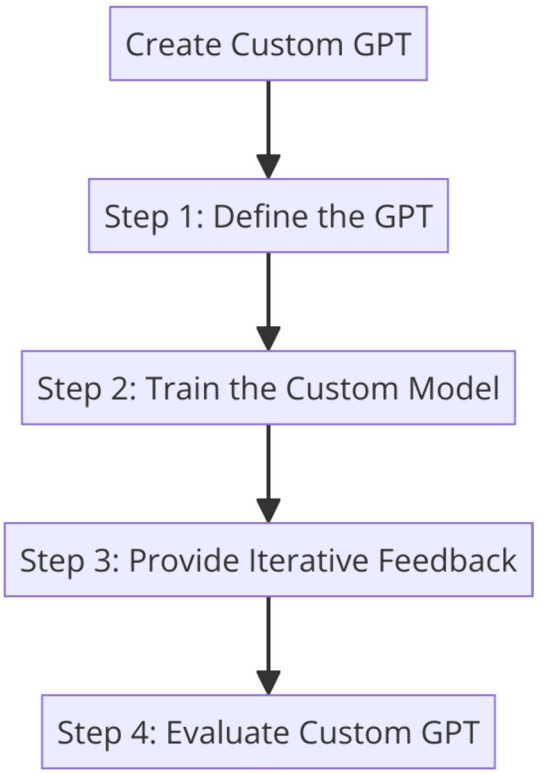

Fig. 1 Customizing a GPT Model Using ChatGPT-4

entered into a fresh chat window without prior chat history in the session to limit additional learning from prior input.

Performance evaluation

Accurate performance was defined as the alignment of ChatGPT responses with guideline recommendations on the surgical management of with GERD [7]. Findings were reported using descriptive statistics, with counts and percentages applied to characterize dichotomous outcomes. Responses that did not provide clinically meaningful advice and guidance conflicting with guideline recommendations were judged not to align with guideline recommendations, demonstrating inaccurate model performance. Two team members evaluated all responses in a blinded fashion to the chatbot model, and no conflicts were generated.

Results

We analyzed a total of 60 cases presented by a hypothetical surgeon and 40 cases presented by a hypothetical patient to evaluate the recommendations provided by the GTS and generic ChatGPT models. The GTS correctly addressed 100% (60/60) of the surgeon's queries and 100% (40/40) of the patient's queries. Conversely, the ChatGPT model exhibited a lower accuracy rate, correctly responding to 66.6% (40/60) of the surgeon's questions and 47.5% (19/40) of the patient's inquiries (see Table 2). Recommendations on the surgical management of GERD generated by the GTS consistently adhered to the SAGES guidelines, whereas those from the generic ChatGPT model did not cite evidence to support its recommendations. No identifiable pattern was observed in the nature of the cases to which the generic ChatGPT-4 model provided incorrect responses.

Discussion

This study evaluated the ability of the GTS, a customized ChatGPT model, to provide recommendations for the surgical management of gastroesophageal reflux disease (GERD) to both surgeons and patients. We observed that the GTS provided very accurate recommendations compared to the generic ChatGPT-4 model. The GTS was 100% accurate in both patient and surgeon inquiries, citing its tailored guidelines. Conversely, the generic ChatGPT-4 model provided inaccurate recommendations frequently to both surgeons and patients without citing evidence to support its guidance. Surgeons and researchers should note that customizing LLMs like ChatGPT-4 could overcome the limitations of generic LLMs for simple topics, as demonstrated by the customized GPT model in this study.

The emergence of LLMs like ChatGPT has created opportunities to access information for patients already seeking clinical advice online [9]. However, the functionality of LLMs can be misconceived [10]. Users may assume that LLMs access the internet in real-time while applying complex algorithms to provide them with the most suitable responses to their query based on these resources. Rather, LLMs rely on complex neural networks developed through an iterative process of input and user feedback [3]. LLMs become sophisticated in their ability to predict the most likely next word in a sequence as opposed to accessing and searching data from its training dataset to reply to a given input. Because of their ability to create language and sentence structure they can appear confident, even when inaccurate. Numerous studies demonstrate that generically trained LLMs provide medical advice with inconsistent reliability. One study [11] found that LLMs occasionally provided incorrect or out-of-date information and cited inappropriate sources, similar to the findings here for the generic ChatGPT-4 model. However, through customization, we achieved a significant improvement in accuracy. Similarly, custom chatbots are outperforming generic LLMs in the setting of urology, and gastroenterology, often with 100% accuracy as reported here [12–14]. Clinicians and researchers may take interest in this approach to customizing GPTs, which avoids the time and resource-intensive nature associated

Table 1 Study prompts

Adult patient	
I'm a surgeon. I have a 45-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40 mg orally twice daily for the last 2 years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?	I am 45 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now. Should I get surgery?
How should I treat my 45-year-old patient with GERD despite PPI therapy	Should I get surgery for heartburn? I'm 45 and take medication
I'm a surgeon. I have a 52-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40 mg orally twice daily for the last 2 years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?	I am 52 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now. Should I get surgery?
How should I treat my 52-year-old patient with GERD despite PPI therapy	Should I get surgery for heartburn? I'm 52 and take medication
I'm a surgeon. I have a 60-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40 mg orally twice daily for the last 2 years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?	I am 60 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now. Should I get surgery?
How should I treat my 60-year-old patient with GERD despite PPI therapy	Should I get surgery for heartburn? I'm 60 and take medication
I'm a surgeon. I have a 67-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40 mg orally twice daily for the last 2 years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?	I am 67 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now. Should I get surgery?
How should I treat my 67-year-old patient with GERD despite PPI therapy	Should I get surgery for heartburn? I'm 67 and take medication
I'm a surgeon. I have a 75-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40 mg orally twice daily for the last 2 years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?	I am 75 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now. Should I get surgery?
How should I treat my 75-year-old patient with GERD despite PPI therapy	Should I get surgery for heartburn? I'm 75 and take medication
I'm a surgeon. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?	
	I am 45 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. Should the surgeon do this with keyhole surgery or with a robot?
Should I operate laparoscopically or robotically on my 45-year-old patient with GERD	Should I get keyhole surgery or robot for my heartburn surgery? I'm 45
I'm a surgeon. I have a 52-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?	
	I am 52 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. Should the surgeon do this with keyhole surgery or with a robot?
Should I operate laparoscopically or robotically on my 52-year-old patient with GERD	Should I get keyhole surgery or robot for my heartburn surgery? I'm 52
I'm a surgeon. I have a 60-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?	
	I am 60 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. Should the surgeon do this with keyhole surgery or with a robot?

Table 1 (continued)

Adult patient	
Should I operate laparoscopically or robotically on my 60-year-old patient with GERD	Should I get keyhole surgery or robot for my heartburn surgery? I'm 60
I'm a surgeon. I have a 67-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?	
	I am 67 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. Should the surgeon do this with keyhole surgery or with a robot?
Should I operate laparoscopically or robotically on my 67-year-old patient with GERD	Should I get keyhole surgery or robot for my heartburn surgery? I'm 67
I'm a surgeon. I have a 75-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?	
	I am 75 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. Should the surgeon do this with keyhole surgery or with a robot?
Should I operate laparoscopically or robotically on my 75-year-old patient with GERD	Should I get keyhole surgery or robot for my heartburn surgery? I'm 75
I'm a surgeon. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like optimal symptom control. Should a partial or complete fundoplication be performed?	
	I am 45 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like full symptom control. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 45-year-old patient? They are worried about symptom control	Should my surgeon do a full wrap or a partial wrap? I'm 45 and I want to stop my heartburn
I'm a surgeon. I have a 52-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like optimal symptom control. Should a partial or complete fundoplication be performed?	
	I am 52 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like full symptom control. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 52-year-old patient? They are worried about symptom control	Should my surgeon do a full wrap or a partial wrap? I'm 52 and I want to stop my heartburn
I'm a surgeon. I have a 60-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like optimal symptom control. Should a partial or complete fundoplication be performed?	
	I am 60 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like full symptom control. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 60-year-old patient? They are worried about symptom control	Should my surgeon do a full wrap or a partial wrap? I'm 60 and I want to stop my heartburn
I'm a surgeon. I have a 67-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like optimal symptom control. Should a partial or complete fundoplication be performed?	
	I am 67 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like full symptom control. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 67-year-old patient? They are worried about symptom control	Should my surgeon do a full wrap or a partial wrap? I'm 67 and I want to stop my heartburn
I'm a surgeon. I have a 75-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like optimal symptom control. Should a partial or complete fundoplication be performed?	

Table 1 (continued)

Adult patient	
	I am 75 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like full symptom control. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 75-year-old patient? They are worried about symptom control	Should my surgeon do a full wrap or a partial wrap? I'm 75 and I want to stop my heartburn
I'm a surgeon. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid dysphagia. Should a partial or complete fundoplication be performed?	
	I am 45 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like to avoid dysphagia. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 45-year-old patient? They are worried about dysphagia	Should my surgeon do a full wrap or a partial wrap? I'm 45 and I'm worried about swallowing after surgery
I'm a surgeon. I have a 52-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid dysphagia. Should a partial or complete fundoplication be performed?	
	I am 52 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like to avoid dysphagia. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 52-year-old patient? They are worried about dysphagia	Should my surgeon do a full wrap or a partial wrap? I'm 52 and I'm worried about swallowing after surgery
I'm a surgeon. I have a 60-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid dysphagia. Should a partial or complete fundoplication be performed?	
	I am 60 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like to avoid dysphagia. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 60-year-old patient? They are worried about dysphagia	Should my surgeon do a full wrap or a partial wrap? I'm 60 and I'm worried about swallowing after surgery
I'm a surgeon. I have a 67-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid dysphagia. Should a partial or complete fundoplication be performed?	
	I am 67 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like to avoid dysphagia. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 67-year-old patient? They are worried about dysphagia	Should my surgeon do a full wrap or a partial wrap? I'm 67 and I'm worried about swallowing after surgery
I'm a surgeon. I have a 75-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid dysphagia. Should a partial or complete fundoplication be performed?	
	I am 75 years old. I have used pantoprazole 40 mg twice a day for the last 2 years. I'm getting more heartburn now and I'm getting surgery. I would like to avoid dysphagia. Should they do a full wrap or a partial wrap?
Should I perform partial or complete fundoplication for my 75-year-old patient? They are worried about dysphagia	Should my surgeon do a full wrap or a partial wrap? I'm 75 and I'm worried about swallowing after surgery
I'm a surgeon. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They want optimum symptom control. Should division of short gastric vessels or no division of short gastric vessels be performed?	
Should I divide the gastric vessels for my 45-year-old patient's fundoplication? They want maximum symptom control	

Table 1 (continued)

Adult patient
I'm a surgeon. I have a 52-year-old patient who is receiving surgical fundoplication for chronic GERD. They want optimum symptom control. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 52-year-old patient's fundoplication? They want maximum symptom control
I'm a surgeon. I have a 60-year-old patient who is receiving surgical fundoplication for chronic GERD. They want optimum symptom control. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 60-year-old patient's fundoplication? They want maximum symptom control
I'm a surgeon. I have a 67-year-old patient who is receiving surgical fundoplication for chronic GERD. They want optimum symptom control. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 67-year-old patient's fundoplication? They want maximum symptom control
I'm a surgeon. I have a 75-year-old patient who is receiving surgical fundoplication for chronic GERD. They want optimum symptom control. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 75-year-old patient's fundoplication? They want maximum symptom control
I'm a surgeon. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid gas bloat and other complications. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 45-year-old patient's fundoplication? They don't want gas bloat
I'm a surgeon. I have a 52-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid gas bloat and other complications. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 52-year-old patient's fundoplication? They don't want gas bloat
I'm a surgeon. I have a 60-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid gas bloat and other complications. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 60-year-old patient's fundoplication? They don't want gas bloat
I'm a surgeon. I have a 67-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid gas bloat and other complications. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 67-year-old patient's fundoplication? They don't want gas bloat
I'm a surgeon. I have a 75-year-old patient who is receiving surgical fundoplication for chronic GERD. They would like to avoid gas bloat and other complications. Should division of short gastric vessels or no division of short gastric vessels be performed?
Should I divide the gastric vessels for my 75-year-old patient's fundoplication? They don't want gas bloat

Table 2 Alignment of recommendations with guidelines generated by the GTS & a generic ChatGPT-4 model for the surgical management of GERD

	Y		ChatGPT-4		SAGES & UEG-EAES
# Aligned w/ guidelines	60/60	35/35	40/60	16/35	
Case #	Surgeon	Patient	Surgery	Patient	
1	Y*	Y	Y	X[!]	Surgery
2	Y	Y	N**	X	Surgery
3	Y	Y	Y	X	Surgery
4	Y	Y	N	X	Surgery
5	Y	Y	Y	X	Surgery
6	Y	Y	N	X	Surgery
7	Y	Y	Y	X	Surgery
8	Y	Y	N	X	Surgery
9	Y	Y	Y	X	Surgery
10	Y	Y	N	X	Surgery
11	Y	Y	Y	Y	CJ[!!]
12	Y	Y	Y	Y	CJ
13	Y	Y	Y	Y	CJ
14	Y	Y	Y	Y	CJ
15	Y	Y	Y	Y	CJ
16	Y	Y	Y	Y	CJ
17	Y	Y	Y	Y	CJ
18	Y	Y	Y	Y	CJ
19		Y	Y	Y	CJ
20	Y	Y	Y	Y	CJ
21	Y	Y	N	N	Complete
22	Y	Y	Y	X	Complete
23	Y	Y	Y	Y	Complete
24	Y	Y	Y	X	Complete
25	Y	Y	Y	Y	Complete
26	Y	Y	Y	X	Complete
27	Y	Y	Y	Y	Complete
28	Y	Y	Y	X	Complete
29	Y	Y	Y	Y	Complete
30	Y	Y	Y	X	Complete
31	Y	Y	Y	Y	Partial
32	Y	Y	Y	X	Partial
33	Y	Y	Y	Y	Partial
34	Y	Y	Y	X	Partial
35	Y	Y	Y	Y	Partial
36	Y	Y	Y	X	Partial
37	Y	Y	Y	Y	Partial
38	Y	Y	Y	X	Partial
39	Y	Y	Y	Y	Partial
40	Y	Y	Y	X	Partial
41	Y		N		Divide
42	Y		Y		Divide
43	Y		N		Divide
44	Y		Y		Divide
45	Y		N		Divide
46	Y		Y		Divide
47	Y		N		Divide
48	Y		Y		Divide
49	Y		N		Divide
50	Y		Y		Divide
51	Y		N		No divide
52	Y		N		No divide
53	Y		N		No divide
54	Y		N		No divide
55	Y		N		No divide
56	Y		N		No divide
57	Y		N		No divide
58	Y		N		No divide
59	Y		Y		No divide
60	Y		N		No divide

*Yes, recommendation aligned with SAGES & UEG-EAES guidelines

**No, recommendation not aligned with SAGES & UEG-EAES guidelines

X Did not provide clinical recommendation

!! Clinical judgement needed based on availability of expertise in robotic versus laparoscopic surgery

with the development of LLMs while potentially enhancing the reliability for clinical decision support [15]. Though ChatGPT restricts access to custom LLMs to its paid users as a closed-source entity that withholds details about its functionality, other open-source LLMs exist and could be similarly customized and integrated into society webpages, clinical workflow via apps or electronic medical health systems via multidisciplinary collaboration with machine learning researchers and data scientists.

Despite their accessibility, caution is warranted. The guidance provided by LLMs is not consistently grounded in clinical evidence, putting patient safety at risk [16]. Numerous studies have highlighted inaccuracies in LLM decision-making and a lack of verifiable resources to support their recommendations. These findings raise ethical concerns regarding the use of LLMs in healthcare settings. In contrast, several studies have justified and recommended the use of LLMs [17] over traditional internet search engines such as Google [18]. Users must be aware of the most updated training data for these LLMs, as they may not be equipped with the

latest updates on treatment guidelines or recommendations depending on the clinical topics and contexts. Additionally, the lack of regulation of LLMs and widespread accessibility make the outputs hard to generalize in terms of accuracy or reliability. The threats to patient safety also include cybersecurity concerns [19], which involve safeguarding patient data and protecting against potential breaches or unauthorized access. Patient data must be handled ethically and in compliance with privacy regulations [20, 21]. Moreover, there is increasing awareness of the potential for bias [22] within LLMs, both in terms of the data they are trained on and the recommendations they generate. Bias in LLMs can manifest in various ways, including disparities in the representation of different demographic groups or medical conditions, which could impact the fairness and equity of clinical advice provided. Therefore, addressing these concerns [23] surrounding cybersecurity, data privacy, and bias is crucial to ensuring the safe and ethical use of LLMs in healthcare. The clinical integration of LLMs must be taken with caution, while acknowledging the potential advantages to integrating LLMs into healthcare practices.

Professional societies, tertiary institutions and hospitals may take interest in exploring the development of online platforms that support customized LLMs accessible via their websites or apps, offering medical advice and addressing common patient inquiries in both inpatient and outpatient settings [24]. Similarly, tailored online tools could be designed for healthcare providers, serving as evidence-based resources for guiding patient management, particularly in complex cases. Institutions [25] stand to benefit significantly from leveraging a customized LLM to optimize communication with patients and automate repetitive tasks. Integrating these clinical tools into healthcare systems could potentially improve patient care and satisfaction while alleviating the workload on healthcare staff. With the proper training dataset, the clinical integration of custom LLMs could potentially lower bias and tailor responses to local populations. Therefore, policymakers and healthcare managers should prioritize the exploration and implementation of these innovative solutions to create positive outcomes for both patients and healthcare professionals. While the potential benefits of these customized models are promising, the proportion of patients using LLMs for health advice is currently unknown. Moreover, an increased emphasis on objective performance evaluation is needed to ensure that LLM advice aligns with the highest quality evidence such as clinical practice guidelines [26], while acknowledging doubt in the setting of more controversial, complex topics.

Limitations exist in this study. Firstly, all cases applied in this pilot study were hypothetical. There is a need for patient-centered, prospective studies to evaluate to efficacy of LLMs in providing clinical advice in a pragmatic context. Moreover, despite the advantages of customizing ChatGPT to provide accurate recommendations based on guidelines, there are significant constraints. One significant limitation is that while we can tailor ChatGPT to our customized model, we are bound by the database on which ChatGPT has been trained. This means that the responses generated by ChatGPT may be influenced by the data it has been exposed to during training, potentially limiting the accuracy of its outputs in healthcare contexts. Moreover, it's essential to acknowledge that ChatGPT was not originally designed or validated for medical use. So, while we may customize it to address medical concerns from surgeons or patients, it lacks the formal validation [26] and regulatory approval required for clinical applications. Establishing a customized ChatGPT tool in real-world medical settings is not advisable without thorough validation studies. Further research is required to assess the reliability, consistency, and safety of using a customized LLM in healthcare practice. Validation studies would be necessary to evaluate its performance in providing accurate and clinically relevant recommendations across a diverse range of medical scenarios. Only through rigorous validation can we establish the trustworthiness and effectiveness of a tailored LLM model as a viable tool for supporting healthcare providers and patients in clinical decision-making.

Conclusion

Clinicians, researchers, and patients may take interest in the ability of a customized version of OpenAI's ChatGPT-4 to significantly improve the accuracy of generating advice for the surgical management of GERD. Customization of the LLM increased patient-focused and provider-focused questions responses substantially. With prior studies illustrating the limitations of LLMs in providing reliably accurate health advice, this approach may be applied to mitigate the time and resource-intensive nature associated with developing de-novo LLMs. The integration of LLMs into clinical practice must be undertaken with the utmost consideration for patient safety, privacy, ethical, and regulatory factors.

Acknowledgements None.

Funding This study received no funding.

Declarations

Disclosures Dr. Danielle S. Walsh is Co-Chair of the Guidelines Committee for Society of Gastrointestinal and Endoscopic Surgeons. Dr. Danielle S. Walsh is a Member of the American College of Surgeons Health Information Technology Committee and the Board of Governors. Dr. Bethany J. Slater is a consultant for Cook Medical and Hologic. Dr. Bethany J. Slater is the Chair of the Guidelines Committee for Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). Dr. Patricia Sylla is a consultant for Safeheal, Ethicon, Stryker and Tissium. Dr. Patricia Sylla is the past president of SAGES.

Drs. Bright Huo, Nana Marfo, Elisa Calabrese, Sunjay Kumar, and Wesley Vosburg have no conflicts of interest to disclose.

Ethics approval Not applicable.

Consent statement Not applicable.

References

1. Dave T, Athaluri SA, Singh S (2023) ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. https://doi.org/10.3389/frai.2023.1169595
2. Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, Pletcher MJ, Lai K (2024) Development of a liver disease-specific large language model chat interface using retrieval augmented generation. Hepatology. https://doi.org/10.1097/hep.0000000000000834
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med. https://doi.org/10.1038/s41591-023-02448-8
4. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, Chang S, Berkowitz S, Finn A, Jahangir E, Scoville E, Reese T, Friedman D, Bastarache J, van der Heijden Y, Wright J, Carter N, Alexander M, Choe J, Wheless L (2023) Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Square. https://doi.org/10.21203/rs.3.rs-2566942/v1
5. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, Wagner SJ, Kather JN (2023) The future landscape of large language models in medicine. Commun Med. https://doi.org/10.1038/s43856-023-00370-1
6. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J (2023) Ethical considerations of using ChatGPT in health care. J Med Int Res. https://doi.org/10.2196/48009
7. Slater BJ, Dirks RC, McKinley SK, Ansari MT, Kohn GP, Thosani N, Qusmeya B, Billeier S, Daly S, Crwaford C, Ehlers AP, Hollands C, Palazzo F, Rodriguez N, Train A, Wassenaar E, Walsh D, Pryor AD, Stefanidis D (2021) SAGES guidelines for the surgical treatment of gastroesophageal reflux (GERD). Surg Endosc. https://doi.org/10.1007/s00464-021-08625-5
8. Markar S, Andreou A, Bonavina L, Florez ID, Huo B, Kontouli KM, Low DE, Mavridis D, Maynard N, Moss A, Pera M, Savarino E, Siersema P, Sifrim D, Watson DI, Zaninotto G, Antoniou SA (2022) UEG and EAES rapid guideline: Update systematic review, network meta-analysis, CINeMA and GRADE assessment, and evidence-informed European recommendations on surgical management of GERD. United European Gastroenterol J 10:983–998. https://doi.org/10.1002/ueg2.12318
9. Ayoub NF, Lee YJ, Grimm D, Divi V (2023) Head-to-head comparison of ChatGPT versus google search for medical knowledge acquisition. Otolaryngol Head Neck Surg. https://doi.org/10.1002/ohn.465
10. Chang IC, Shih YS, Kuo KM (2022) Why would you use medical chatbots? Interview and survey. Int J Med Inform. https://doi.org/10.1016/j.ijmedinf.2022.104827
11. Cung M, Sosa B, Yang HS, McDonald MM, Matthews BG, Vlug AG, Imel EA, Wein MN, Stein EM, Greenblatt MB (2024) The performance of AI chatbot large language models to address skeletal biology and bone health queries. J Bone Miner Res. https://doi.org/10.1093/jbmr/zjad007
12. Khene ZE, Bigot P, Mathieu R, Rouprêt M, Bensalah K (2024) Development of a personalized chat model based on the European Association of Urology Oncology Guidelines: harnessing the power of generative artificial intelligence in clinical practice. Eur Urol Oncol. https://doi.org/10.1016/j.euo.2023.06.009
13. Simsek C, Madaria E, Ebigbo A, Vanek P, Elshaarawy O, Voiosu A, Antonelli G, Turro R, Gisbert J, Nyssen O, Messmann H, Cesare H, Jalan R, Demir H, Tinaz B, Erol M (2024) Gastrogpt: development and controlled testing of a proof-of concept customized clinical language model. Lancet. https://doi.org/10.2139/ssrn.4718227
14. Tariq R, Voth E, Khanna S (2024) Integrating clinical guidelines with ChatGPT-4 enhances its' skills. Mayo Clin Proc. https://doi.org/10.1016/j.mcpdig.2024.02.004
15. Wang Y, Visweswaran S, Kapoor S, Kooragayalu S, Wu X (2024) ChatGPT-CARE: a superior decision support tool enhancing ChatGPT with clinical practice guidelines. medRxiv. https://doi.org/10.1101/2023.08.09.23293890
16. Haupt CE, Marks M (2023) AI-generated medical advice - GPT and beyond. JAMA. https://doi.org/10.1001/jama.2023.5321
17. Henson JB, Glissen Brown JR, Lee JP, Patel A, Leiman DA (2023) Evaluation of the potential utility of an artificial intelligence chatbot in gastroesophageal reflux disease management. Am J Gastroenterol. https://doi.org/10.14309/ajg.0000000000002397
18. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM (2023) Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res. https://doi.org/10.2196/47479
19. Parviainen J, Rantala J (2022) Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. Med Health Care Philos. https://doi.org/10.1007/s11019-021-10049-w
20. Hasal M, Nowaková J, Ahmed Saghair K, Abdulla H, Snášel V, Ogiela L (2021) Chatbots: security, privacy, data protection, and social aspects. Concurr Comput Pract Exp. https://doi.org/10.1002/cpe.6426
21. Hacker P, Engel A, Mauer M (2023) Regulating ChatGPT and other large generative AI models. arXiv. https://doi.org/10.1145/3593013.3594067
22. McGreevey JD, Hanson CW, Koppel R (2020) Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. JAMA. https://doi.org/10.1001/jama.2020.2724
23. Chow JCL, Sanders L, Li K (2023) Impact of ChatGPT on medical chatbots as a disruptive technology. Front Art Intell. https://doi.org/10.3389/frai.2023.1166014
24. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E (2023) Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? Diagnostics. https://doi.org/10.3390/diagnostics13111950
25. Javaid M, Haleem A, Singh RP (2023) ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval. https://doi.org/10.1016/j.tbench.2023.100105
26. Ritchie JB, Frey LJ, Lamy JB, Bellcross C, Morrison H, Schiffman JD, Welch BM (2022) Automated clinical practice guideline recommendations for hereditary cancer risk using chatbots and ontologies: system description. JMIR Cancer. https://doi.org/10.2196/29289