




# The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease

Bright Huo<sup>1</sup> · Elisa Calabrese<sup>2</sup> · Patricia Sylla<sup>3</sup> · Sunjay Kumar<sup>4</sup> · Romeo C. Ignacio<sup>5</sup> · Rodolfo Oviedo<sup>6,7,8</sup> · Imran Hassan<sup>9</sup> · Bethany J. Slater<sup>10</sup> · Andreas Kaiser<sup>11</sup> · Danielle S. Walsh<sup>12</sup> · Wesley Vosburg<sup>13</sup> 

Received: 12 March 2024 / Accepted: 21 March 2024 / Published online: 17 April 2024  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

## Abstract

**Background** Large language model (LLM)-linked chatbots may be an efficient source of clinical recommendations for healthcare providers and patients. This study evaluated the performance of LLM-linked chatbots in providing recommendations for the surgical management of gastroesophageal reflux disease (GERD).

**Methods** Nine patient cases were created based on key questions addressed by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) guidelines for the surgical treatment of GERD. ChatGPT-3.5, ChatGPT-4, Copilot, Google Bard, and Perplexity AI were queried on November 16th, 2023, for recommendations regarding the surgical management of GERD. Accurate chatbot performance was defined as the number of responses aligning with SAGES guideline recommendations. Outcomes were reported with counts and percentages.

**Results** Surgeons were given accurate recommendations for the surgical management of GERD in an adult patient for 5/7 (71.4%) KQs by ChatGPT-4, 3/7 (42.9%) KQs by Copilot, 6/7 (85.7%) KQs by Google Bard, and 3/7 (42.9%) KQs by Perplexity according to the SAGES guidelines. Patients were given accurate recommendations for 3/5 (60.0%) KQs by ChatGPT-4, 2/5 (40.0%) KQs by Copilot, 4/5 (80.0%) KQs by Google Bard, and 1/5 (20.0%) KQs by Perplexity, respectively. In a pediatric patient, surgeons were given accurate recommendations for 2/3 (66.7%) KQs by ChatGPT-4, 3/3 (100.0%) KQs by Copilot, 3/3 (100.0%) KQs by Google Bard, and 2/3 (66.7%) KQs by Perplexity. Patients were given appropriate guidance for 2/2 (100.0%) KQs by ChatGPT-4, 2/2 (100.0%) KQs by Copilot, 1/2 (50.0%) KQs by Google Bard, and 1/2 (50.0%) KQs by Perplexity.

**Conclusions** Gastrointestinal surgeons, gastroenterologists, and patients should recognize both the promise and pitfalls of LLM's when utilized for advice on surgical management of GERD. Additional training of LLM's using evidence-based health information is needed.

**Keywords** GERD · Surgery · ChatGPT · Generative artificial intelligence · Natural language processing · Large language models · Guidelines

Web-based large language models (LLMs) are a subsets of artificial intelligence and have become popular, with ChatGPT reaching over 100 million users shortly after its release [1]. These artificial intelligence platforms undergo multi-stage training using articles, books, and other online content to generate conversational, human-like responses to user queries [2]. LLMs iteratively learn language through word associations during this process to recognize, interpret, and generate text without fine-tuning [2]. Due to their public

accessibility and convenient user interface [3, 4], there is significant interest in the ability of LLMs to provide the user with recommendations for healthcare-related queries [5–7]. Up to 90% of Internet users including both patients and clinicians search for health-related information online [8–10], as it is immediate, convenient, and generally free. Moreover, up to 80% of Internet users feel that the online health information which they retrieve is reliable [11].

However, the responses generated by LLM's are not verified by health professionals, leading to concerns about the accuracy and safety of chatbot medical advice.[9] The provision of inaccurate clinical recommendations by chatbots has

Extended author information available on the last page of the article

the potential to negatively impact patient safety [12]. While most chatbots provide a disclaimer that the responses should not be taken as medical advice, the healthcare community has an obligation to study and report the performance of these tools on behalf of our patients, especially while more rigorous standards for assessment are still in development [13]. The accuracy of clinical recommendations provided by LLM-linked chatbots has health implications for patients with common medical problems.

Gastroesophageal reflux disease (GERD) affects 18.1–27.8% of North Americans [14]. It has been reported that 93% of ChatGPT-derived recommendations for the management of GERD is appropriate based on expert physician opinion [15]. However, GERD can be managed with various medical and surgical options, increasing the difficulty of making treatment decisions [16]. Surgical decision-making in the treatment of GERD is especially multi-factorial [17], necessitating various technical considerations [16]. Patient factors, response to treatment, complicated diagnostic studies, and patient-tailored assessments are all incorporated into successful strategies. Thus, gastrointestinal surgeons, gastroenterologists, primary care providers, patients, and researchers would benefit from a structured investigation of the ability of chatbots to provide accurate treatment advice for GERD.

Given the short timeframe in which LLM-linked chatbots have been sensationalized, the use of objective measures of clinical performance among chatbots remain early in development and validation. High-quality Chatbot Assessment Studies must report transparent, reproducible methodology to facilitate the interpretation of study findings by readers. In the absence of formal evaluation tools for Chatbot Assessment Studies, the use of standardized patient cases with expert input and assessment based on high-quality evidence would facilitate the evaluation of chatbot performance in providing clinical recommendations. Thus, the aim of this study was to assess the performance of LLM-linked chatbots in providing recommendations for the surgical management of GERD using recently published SAGES Guidelines as an objective measure of chatbot performance [16].

## Materials and methods

### Study objectives

The primary objective of this study was to assess whether LLM-linked chatbots could provide accurate recommendations for the surgical management of GERD to both patients and surgeons based on their alignment with SAGES guideline recommendations. Secondary objectives were to evaluate whether LLM-linked chatbots could provide accurate ratings of the certainty of the evidence based on their alignment

with SAGES guideline ratings, as well as to identify whether chatbots would provide incongruent recommendations to patients and surgeons. Evidence cited by chatbots to support their recommendations was also explored.

### Query strategy

Hypothetical adult and pediatric patient cases and prompts were based on key questions (KQs) from the SAGES guidelines for the surgical treatment of gastroesophageal reflux (GERD) [16]. The patient cases were constructed to reflect the population, intervention, and comparator addressed by clinical recommendations for each KQ. This information was combined with input from expert general and foregut surgeons to develop clinical questions that were phrased with appropriate medical terminology for surgeon inquiries, while KQs for patient inquiries were worded using lay terminology. The pediatric patient case prompts were phrased such that a parent or caregiver was asking the chatbot for recommendations for their child.

On November 16th, 2023, prompts were tested across LLM-linked chatbots including ChatGPT-4 (GPT-4-0613), Copilot (formerly Bing Chat), Google Bard, and Perplexity AI. These LLMs were chosen among the most frequently assessed chatbots for clinical application based on an internal scoping review. Bing Chat was queried prior to its full rebranding to Copilot on December 1st. Copilot (and Bing Chat on the date of query) is built on OpenAI's GPT-4 and DALL-E 3. Copilot was set in the "More Precise" mode. This was done to identify generic chatbot responses or responses that did not provide meaningful information. Google Bard uses an experimental model named PaLM 2 that was last updated on the day of the search query. Perplexity AI accessed their fine-tuned version of OpenAI's GPT-3.5 using the "co-pilot" mode for all queries. During this time, follow-up prompts were trialed to bypass obstructive responses such as legal disclaimers, which would otherwise dilute the meaningful information obtained. The patient case and KQs were used in standardized prompts to query LLM-linked chatbots for clinical recommendations for surgeons (Table 1) and for patients (Table 2). Follow-up prompts were created, and specific scenarios for their use were defined a priori (Table 3). Neither prompts nor follow-up prompts contained any reference to major surgical societies, organizations, or countries to mitigate bias. All prompts were reviewed by a second team member for grammatical correctness, as well as appropriateness for the study. The English language was used for all prompts.

ChatGPT-4, Copilot, Google Bard, and Perplexity AI were queried on November 16th, 2023, from a computer server in Hamilton, Ontario, Canada. A hotspot program was used to access Google Bard from the USA, as it is not yet accessible in Canada. All LLMs were freely accessible

**Table 1** Clinician prompts

Adult patient	Pediatric patient	Obstructive responses	Specific follow-up
<p>I'm a surgeon in North America. I have a 45-year-old patient. They have been managing their chronic gastroesophageal reflux disease using pantoprazole 40mg orally twice daily for the last two years. They have gradually experienced a worsening of their symptoms. Should they continue with medical or surgical management?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 12-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>Medical management must be optimized. Depends on whether the patient is concerned about long-term reflux control</p> <p>Depends on whether dysphagia is a current symptom for the patient</p> <p>Consult alternative specialist</p> <p>A full work-up must be completed, various tests mentioned</p> <p>It depends on the severity of their symptoms</p> <p>It depends on their comorbidity status</p>	<p>Medical management is optimized. Please make a recommendation</p> <p>The patient prefers long-term reflux control. Please make a recommendation</p> <p>The patient is not experiencing dysphagia. Please make a recommendation</p> <p>The patient has already been reviewed by that specialist</p> <p>A full work-up has been completed</p> <p>Their symptoms are severe</p> <p>They are otherwise healthy</p>
<p>I'm a surgeon in North America. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They are very bothered by having to take a PPI. Should this be performed laparoscopically or robotically?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 12-year-old patient who is receiving surgical fundoplication for chronic GERD. Should this be performed laparoscopically or robotically?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about the laparoscopic vs robotic approach</p>	<p>The surgeon is comfortable with either approach. Please make a recommendation</p>
<p>I'm a surgeon in North America. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They are worried about the effectiveness of surgery, the need for a second operation, and postoperative complications. Should this be performed laparoscopically or robotically?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. They are worried about the effectiveness of surgery, the need for a second operation, and postoperative complications. Should this be performed laparoscopically or robotically?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about the laparoscopic vs robotic approach</p>	<p>The surgeon is comfortable with either approach. Please make a recommendation</p> <p>They have occasional bloating or belching</p> <p>Their esophageal motility is unknown</p>
<p>I'm a surgeon in North America. 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?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 45-year-old patient who is receiving surgical fundoplication for chronic GERD. Should a partial or complete fundoplication be performed?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about performing partial vs complete repair</p> <p>Depends on their symptoms of bloating or belching</p> <p>Depends on their esophageal motility</p>	<p>The surgeon is comfortable with either approach. Please make a recommendation</p> <p>They have occasional bloating or belching</p> <p>Their esophageal motility is unknown</p>
<p>I'm a surgeon in North America. 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?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. 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?</p> <p>What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about performing partial vs complete repair</p> <p>Depends on their symptoms of bloating or belching</p> <p>Depends on their esophageal motility</p>	<p>The surgeon is comfortable with either approach. Please make a recommendation</p> <p>They have occasional bloating or belching</p> <p>Their esophageal motility is unknown</p>

**Table 1** (continued)

Adult patient	Pediatric patient	Obstructive responses	Specific follow-up
<p>I'm a surgeon in North America. 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? What is the certainty of this recommendation based on the quality of the evidence?</p> <p>I'm a surgeon in North America. 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? What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 12-year-old patient who is receiving surgical fundoplication for chronic GERD. Should minimal dissection or maximal dissection be used in pediatric patients with GERD? What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about performing a good repair with or without taking the gastric vessels</p> <p>They should speak to the surgeon about performing a good repair with or without taking the gastric vessels</p>	<p>The surgeon is comfortable with their repair using either approach. Please make a recommendation</p> <p>The surgeon is comfortable with their repair using either approach. Please make a recommendation</p>
<p>I'm a surgeon in North America. I have a 12-year-old patient who is receiving surgical fundoplication for chronic GERD. Should minimal dissection or maximal dissection be used in pediatric patients with GERD? What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>I'm a surgeon in North America. I have a 12-year-old patient who is receiving surgical fundoplication for chronic GERD. Should minimal dissection or maximal dissection be used in pediatric patients with GERD? What is the certainty of this recommendation based on the quality of the evidence?</p>	<p>They should speak to the surgeon about using minimal versus maximal dissection Depends on patient's goals regarding long-term control versus complications</p>	<p>The surgeon is comfortable with their repair using either approach. Please make a recommendation</p> <p>The patient wants long-term control and wants to avoid complications</p>

**Table 2** Patient prompts

Patients	Obstructive responses	Specific follow-up
Should I get surgery for heartburn?		
I am 45 years old. I have used pantoprazole 40mg twice a day for the last two years. I'm getting more heartburn now. Should I get surgery?	Medical management must be optimized Depends on whether you are concerned about long-term reflux control	I have (My child has) tried all the medications. What will give me (my child) the best control of my heartburn? I want (My child wants) long-term reflux control. What will give me (my child) the best control of my heartburn?
I am 45 years old. I have used pantoprazole 40mg twice a day for the last two years. I'm getting more heartburn now and I'm getting surgery. Should the surgeon do this with keyhole surgery or with a robot?	You should speak to your surgeon about keyhole surgery vs robot	The surgeon said that they can use either approach. What will give me the best control of my heartburn?
I am 45 years old. I have used pantoprazole 40mg twice a day for the last two years. I'm getting more heartburn now and I'm getting surgery. Should they do a full wrap or a partial wrap?	You should speak to your surgeon about performing partial vs complete wrap	The surgeon said that they can use either approach. What will give me the best control of my heartburn?
N/A. Short gastric vessel question not applicable to patients	–	–
N/A. Minimal vs maximal dissection question not applicable to patients	–	–

**Table 3** Follow-up prompts

General obstructive chatbot responses	Follow-up prompts—surgeons	Follow-up prompts—patients
<i>I am not a doctor</i>	<ol style="list-style-type: none"> <li>1. I am a surgeon. Repeat prompt</li> <li>2. I acknowledge this disclaimer. Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. I acknowledge this disclaimer. Please make a recommendation</li> <li>2. I have spoken to a surgeon. Please make a recommendation</li> </ol>
<i>Describes pros and cons of proceeding with both the intervention and comparator</i>	<ol style="list-style-type: none"> <li>1. I'm a surgeon. Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. I have spoken to a surgeon. Please make a recommendation</li> </ol>
<i>Lists investigations to be completed</i>	<ol style="list-style-type: none"> <li>2. A full work-up has been completed. Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. I have a diagnosis. The tests are done. Please make a recommendation</li> </ol>
<i>Consider a surgical consultation</i>	<ol style="list-style-type: none"> <li>2. I'm a surgeon. Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. I have spoken to a surgeon. Please make a recommendation</li> </ol>
<i>Consider involving gastroenterology, physicians from other specialties, other staff, etc</i>	<ol style="list-style-type: none"> <li>1. They have been reviewed by gastroenterology. Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. I have seen a gastroenterologist. Please make a recommendation</li> </ol>
<i>Choice depends on patient preference and/or provider comfort</i>	<ol style="list-style-type: none"> <li>1. The surgeon is comfortable with (either option). Please make a recommendation</li> <li>2. The patient is agreeable to (either option). Please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. The surgeon is comfortable with either approach. Please make a recommendation</li> <li>2. The surgeon is comfortable with either approach. What will give me the best control of my heartburn?</li> </ol>
<i>*No guideline or evidence synthesis is cited*</i>	<ol style="list-style-type: none"> <li>1. What do the guidelines recommend?</li> <li>2. Based on what the guidelines state, please make a recommendation</li> </ol>	<ol style="list-style-type: none"> <li>1. What option will give me the best control of my heartburn?</li> </ol>
<i>It depends on the severity of their symptoms</i>	<ol style="list-style-type: none"> <li>1. Their symptoms are severe</li> </ol>	<ol style="list-style-type: none"> <li>1. My symptoms are severe</li> </ol>

with the exception of ChatGPT-4, with a rough cost of \$20 USD monthly at the time of writing. All chatbots were queried by two different study team members using the same prompts and follow-up prompts to ensure the consistency of recommendations made by the chatbots (Supplementary Appendix 1). All prompts were entered into a fresh chat window without prior history in the session. Prompts for surgeon inquiries were entered sequentially in separate chat windows from those utilized for patient inquiries. For surgeons, all prompts began with “I am a surgeon” to prime the chatbots. For patients, all prompts employed layperson terminology, such as “Should I receive surgery for heartburn?” Specific responses for which the use of follow-up prompts was indicated were identified a priori during the prompt testing phase. For surgeons, these included but were not limited to medical disclaimers that the chatbot is not a doctor and/or could not provide medical recommendations, being told to seek a surgical consultation, being told that the patient should trial more medications and lifestyle modifications, and being told to seek the opinions of physicians of other specialties, and other health professionals (Table 3).

### Performance evaluation & response classification

Accurate performance was defined as the alignment of LLM advice with current SAGES guideline recommendations for adult and pediatric patients with GERD. Additionally, we evaluated whether LLM-linked chatbots could accurately cite the certainty of the evidence based on the alignment of chatbot responses with SAGES guideline statements for the

certainty of the evidence. The certainty of the evidence characterizes the strength of the evidence used to make guideline recommendations. A data collection form was developed to collate prompts and response data. Descriptive statistics were used to report dichotomous outcomes including counts and percentages. Dichotomous outcomes included whether responses to prompts aligned with guideline recommendations or not. Responses that were judged not to align with guideline recommendations included those providing guidance conflicting with SAGES guideline recommendations, those without meaningful answers, and those that did not make a recommendation for or against an intervention or comparator. Responses were judged to be successful if they gave guidance that was concordant with SAGES guideline recommendations or gave a recommendation that it was “reasonable” or “appropriate” to proceed with a given intervention or comparison aligned with SAGES recommendations. Surgeon recommendations were compared to patient recommendations to identify the presence of incongruent guidance. When comparing information given to surgeons and patients, recommendations were classified as discordant when recommendations given to surgeons contradicted those given to patients and when no meaningful guidance was given to one group while a recommendation was given to the other. Chatbot guidance that was indifferent for either the intervention or comparator was considered to be correct if corresponding guideline recommendations also did not recommend one intervention over another in any situation. Similarly, certainty of the evidence was judged to be accurate based on the alignment of LLM ratings of certainty

of evidence with SAGES guideline ratings for certainty of evidence for each recommendation. Two team members evaluated responses in a blinded fashion so that they could not identify the chatbots that produced any given response. Conflicts were resolved using a synchronous session. A third expert general surgeon team member was available to resolve conflicts as needed. All researchers were trained on response evaluation through exposure to the above criteria and three pilot questions. Evidence cited by chatbots to support their recommendations was described in narrative form.

## Results

Accurate recommendations for the surgical management of GERD in an adult were provided for 5/7 (71.4%) KQs by ChatGPT-4, 3/7 (42.9%) KQs by Copilot, 6/7 (85.7%) KQs by Google Bard, and 3/7 (42.9%) KQs by Perplexity according to the SAGES guidelines (Table 4). The certainty of the evidence was appropriately provided for 4/7 (57.1%) KQs by ChatGPT-4, 2/7 (28.6%) KQs by Copilot, 3/7 (42.9%) KQs by Google Bard, and 0/7 (0.0%) KQs by Perplexity based on guideline recommendations (Table 4). Patient recommendations for an adult were appropriately given for 3/5 (60.0%) KQs by ChatGPT-4, 2/5 (40.0%) KQs by Copilot, 4/5 (80.0%) KQs by Google Bard, and 1/5 (20.0%) KQs by Perplexity, respectively (Table 4).

ChatGPT-4 gave no clinically meaningful recommendations when asked for a recommendation to proceed with laparoscopic versus robotic surgery for a patient who was bothered by their PPI use. Based on the SAGES guidelines, no chatbot provided the correct recommendations for robotic versus laparoscopic fundoplication for an adult concerned about the effectiveness of surgery, the need for reoperation, and postoperative complications.

Accurate recommendations for the surgical management of GERD in a child were provided for 2/3 (66.7%) KQs by ChatGPT-4, 3/3 (100.0%) KQs by Copilot, 3/3 (100.0%) KQs by Google Bard, and 2/3 (66.7%) KQs by Perplexity according to the SAGES guidelines. The certainty of evidence was appropriately provided for 0/3 (100.0%) KQs by ChatGPT-4, 0/3 (100.0%) KQs by Copilot, 2/3 (66.7%) KQs by Google Bard, and 0/3 (0.0%) KQs by Perplexity based on guideline recommendations (Table 4). Recommendations for a pediatric patient were appropriately given for 2/2 (100.0%) KQs by ChatGPT-4, 2/2 (100.0%) KQs by Copilot, 1/2 (50.0%) KQs by Google Bard, and 1/2 (50.0%) KQs by Perplexity based on SAGES guidelines (Table 4). All chatbots responded with clinically meaningful recommendations. No chatbot was able to appropriately rate the certainty of the recommendation for minimal vs maximal dissection based on the quality of the evidence for a child receiving surgical fundoplication for refractory GERD (Table 5).

**Table 4** Alignment of Chatbot Responses with SAGES Guideline Recommendations for Adults

Chatbot	Chat-GPT-4		Bing chat		Google bard		Perplexity		SAGES
	S	P	S	P	S	P	S	P	
Surgeons (S) or Patient (P) Inquiries	S	P	S	P	S	P	S	P	
Advice Aligned with SAGES	5/7	3/5	3/7	2/5	6/7	4/5	3/7	1/5	–
Certainty Aligned with SAGES	4/7		2/7		3/7		0/7		–
Case	49 y/o on pantoprazole 40mg po BID×2 years for refractory GERD, with worsening reflux symptoms								–
# Clinical Questions									–
1 Medication vs surgery? Certainty?	Y	Y	Y	Y	Y	Y	Y	Y	Surgery Very low
2 Robotics vs laparoscopy—PPI? Certainty?	Y	X*	N	N	Y	Y	N	N	Robot Low
3 Robotics vs laparoscopy—complications? Certainty?	N	N	N	N	N	N	N	N	Laparoscopy Low
4 Complete vs partial repair—symptoms? Certainty?	N	Y	Y	N	Y	Y	N	N	Complete Low
5 Complete vs partial repair—dysphagia? Certainty?	Y	Y	Y	Y	Y	Y	Y	N	Partial Low
6 Dissection of gastric vessels—symptoms? Certainty?	Y		N		Y		Y		Divide Very low
7 Dissection of gastric vessels—gas bloat? Certainty?	Y		N		Y		Y		Don't divide Very low

\*X = Did not provide clinically meaningful recommendation



**Table 5** Alignment of chatbot responses with SAGES guideline recommendations for children

Chatbot	Chat-GPT-4		Bing chat		Google bard		Perplexity		SAGES
	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	
Surgeons (S) or Patient (P) Inquiries	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	<i>S</i>	<i>P</i>	
Advice Aligned with SAGES	2/3	2/2	3/3	2/2	3/3	1/2	2/3	1/2	–
Certainty Aligned with SAGES	0/3		0/3		2/3		0/3		–
Case	12 y/o on pantoprazole 40mg po BID × 2 years for refractory GERD, with worsening reflux symptoms								
# Clinical Questions									
1 Robotics vs laparoscopy?	Y	Y	Y	Y	Y	Y	N	N	Preference
Certainty?	N		N		Y		N		Low
2 Complete vs partial repair?	Y	Y	Y	Y	Y	N	Y	Y	Preference
Certainty?	N		N		N		N		Low
3 Minimal vs maximal dissection?	N		Y		Y		Y		Min
Certainty?	N		N		N		N		Very low

\*X = Did not provide clinically meaningful recommendation

ChatGPT-4 cited recommendations from the American College of Gastroenterology (ACG) 2022, SAGES 2021, and the United European Gastroenterology (UEG)/European Association of Endoscopic Surgery (EAES) 2021 guidelines. Copilot cited SAGES guidelines from 2021. Google Bard cited guidance from the ACG 2021, American College of Physicians (ACP) 2015, SAGES 2021, and joint recommendations from the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition (NASPGHAN) and the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) guidelines. Perplexity AI cited recommendations from the SAGES 2021 guidelines.

## Discussion

This study evaluated the ability of various LLM-linked chatbots to provide recommendations for the surgical management of GERD based on guidelines published by SAGES [16]. We observed that LLM-linked chatbots provided recommendations with inconsistent accuracy. These LLMs also provided discrepant responses for both physician and patient inquiries. We found that Google Bard was most accurate in providing recommendations for both physicians and patients when compared to ChatGPT 4.0, Copilot, and Perplexity. ChatGPT 4.0 followed closely behind Google Bard in the accuracy of information. ChatGPT-4 provided a marginally higher accuracy in its certainty than Google Bard, with Perplexity performing the worst in this domain. However, none of the chatbots provided correct guidance for all clinical questions based on SAGES guideline recommendations. We also found that chatbot-derived ratings for the certainty of the evidence underlying their recommendations were often inaccurate based on ratings from the SAGES guidelines. Though there is promise in the clinical application of

chatbots for patient and physician recommendations, significant improvements must be made to optimize safety for both adult and pediatric patients.

Machine learning has been used to identify patients at risk for acute appendicitis and choledocholithiasis in children [18, 19]. In adults, machine learning has been used to predict patients at risk for GERD following bariatric surgery [20], as well as classify the severity of GERD using endoscopic images [21]. However, the accuracy of LLM-linked chatbots in providing clinical recommendations for patients and clinicians for the management of GERD is not well characterized despite a growing population searching for health advice online [8]. Our findings suggest that chatbots perform differently when providing clinical advice for adult patients compared to pediatric patients. Google Bard and ChatGPT-4 answered the highest proportion of key questions correctly for adults, but Copilot and Google Bard performed the best for cases relating to children. It is noteworthy that prompts were initially tested in Google Bard prior to other LLMs during the development of standardized prompts. Utilizing the tool itself to test the prompts has the potential to significantly impact model output as the tool learns during the training phase [22]. This “pretraining” process may explain the superior performance of Google Bard to ChatGPT demonstrated here. While Rahsepar et al. reported that the converse was true when evaluating lung cancer screening recommendations provided by these LLMs [22], computer scientists recognize that reinforcement learning, as described in creating the scenarios, can both strength and weaken model behaviors [23]. The unpredictability of these tools remains a challenge to be overcome before wide adoption in clinical application for patient care.

No LLM provided consistently accurate recommendations for all key questions based on SAGES guideline recommendations in this study. Similarly, Henson and colleagues



interrogated ChatGPT for guidance surrounding the diagnosis and management of GERD and found that 29% of questions were answered with complete appropriateness, while 62.3% were considered mostly appropriate [15]. In both studies, inappropriate or incorrect clinical guidance would have been provided to both surgeons and patients. Even the highest-performing model for accurate recommendations, Google Bard, provided guidance that conflicted with SAGES guideline recommendations. A significant limitation to LLMs is that they are susceptible to experiencing hallucinations—that is, generating confident answers that may be false, or are not justified by their training data [24]. As many of these models are freely accessible online, this poses a significant risk to patient safety. Many online news reports state that ChatGPT and other LLMs may provide comparable health management to a physician, largely based on studies showing that they can pass licensing examinations [25, 26], or even respond to patient inquiries with greater empathy than clinicians [5]. However, it is essential to recognize that these chatbots predict the next word in a phrase based on the language that they have learned their training datasets [2]. In this context, these models are not synthesizing and interpreting evidence to provide clinical recommendations such as the approach used in clinical practice guidelines. Our study is the first 0-shot evaluation to highlight this in the clinical context, as the ability of LLMs to rate the certainty of the evidence supporting their clinical recommendations was poor.

Gastrointestinal surgeons that perform anti-reflux and foregut surgery, and other clinicians such as family physicians, gastroenterologists, and allied health professionals involved in the treatment of GERD, as well as patients with GERD should be aware of the limitations of LLM-linked chatbots in providing clinical advice. Despite their increased accessibility, these chatbots do not synthesize evidence directly, must often be prompted to provide citations, and are demonstrated to provide inaccurate information. The application of current LLM-linked chatbots in the clinical setting may negatively impact patient care. Prior to the entrance of LLMs into the mainstream, just under half of adults were searching the internet for health information or advice [8, 11], including Google or Wikipedia [27, 28]. Moreover, 80% of patients perceive these online resources to be reliable [11]. Few comparisons have been made between LLMs and traditional online sources. One study reported that health advice for postoperative otolaryngology care generated by ChatGPT scored lower in understandability, actionability, and procedure-specific content than Google [29]. However, different prompts were used to search ChatGPT versus Google, which clouds the interpretation of their findings. In contrast, Hristidis and colleagues found that ChatGPT generated more relevant responses compared to Google for health information

related to dementia [30]. Furthermore, the ability of LLMs to conveniently provide a single resource to synthesize online information will only increase the amount of internet users. Without regulation and quality improvement, the clinical advice from LLMs may impact the ability of patients to understand the treatment plans recommended for them, with the potential to negatively impact their care. Policymakers and hospital managers should take note that LLMs are currently not able to reliably provide accurate recommendations for patients. However, as these models improve, we will likely see their gradual integration into health systems used in the hospital setting. Particularly, the use of institutional data to train closed, inaccessible models to generate tailored patient recommendations based on local outcomes is a key area for future research. Mahajan and colleagues successfully trained a machine learning model using hospital network data to develop a surgical risk prediction tool [31]. Furthermore, these models could be trained to develop a publicly accessible LLM that summarizes societal recommendations as a central resource. Still, this innovative movement must be done with the utmost regard for patient safety, balancing their potential to positively transform patient care and their shortcomings.

Limitations exist in this study. Though high-quality guidelines were used as an objective measure of performance, the quality of currently available primary data limits the certainty of the evidence for many guideline recommendations. Certain surgeon prompts such as minimal versus maximal dissection of short gastric vessels in adult patients could not be answered due to the lack of literature available to inform guideline recommendations. Additionally, these LLM-linked chatbots are not trained specifically for medical application. Most chatbots are closed/proprietary models, and little is known about their functionality. Generally, LLMs are also limited by the information learned from their training dataset which may further impact their performance in a clinical setting. Notably, LLM-linked chatbots are dynamically improving and the results of this study apply to the current state of machine learning. Finally, prompts were not generated by patients, and the results of this study must be interpreted accordingly. The strength of this study is its rigorous methodology including its use of a transparent testing phase, standardized prompts, and an objective measure of LLM performance. While reporting guidelines are in development [13], it is imperative that future Chatbot Assessment Studies adhere to robust methodology and transparent reporting standards. Emphasis must be placed on the development of “open” or accessible LLMs that are trained using clinical datasets. The potential for the use of local datasets to develop LLMs capable of supporting surgical decision-making based on institution-tailored outcome data cannot be understated.

## Conclusion

LLM-linked chatbots are a promising technology within the field of artificial intelligence. Though their widespread accessibility and simple linguistic abilities position them well to support patients and providers with health recommendations, they currently perform surgical decision-making with inconsistent accuracy. Gastrointestinal surgeons, gastroenterologists, and other healthcare professionals involved in the management of patients with GERD must be aware of the potential for future patients to present to their care following the use of LLMs for health recommendations, as well as their current limitations. Policymakers and hospital managers must recognize the potential of LLMs to greatly improve patient care in the clinical setting and be aware of their gradual integration into health systems and applications, but these advancements must be conducted with the utmost consideration for patient safety.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00464-024-10807-w>.

**Acknowledgements** The authors would like to thank the SAGES Guideline Committee for their expert guidance in the development of this manuscript.

**Funding** This study received no funding.

## Declarations

**Disclosures** Walsh is Co-Chair of the Guidelines Committee for Society of Gastrointestinal and Endoscopic Surgeons. Walsh is a Member of the American College of Surgeons Health Information Technology Committee and the Board of Governors. Slater is a consultant for Cook Medical and Hologic. Slater is the Chair of the Guidelines Committee for Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). Sylla is a consultant for Safeheal, Ethicon, Stryker and Tissium. Sylla is the president of SAGES. Huo, Calabrese, Kumar, Ignacio, Oviedo, Hassan, Kaiser, and Vosburg have no conflicts of interest to disclose.

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

## References


- Meyer JG, Urbanowicz RJ, Martin PCN, O'Connor K, Li R, Peng PC, Bright TJ, Tatonetti N, Won KJ, Gonzalez-Hernandez G, Moore JH (2023) ChatGPT and large language models in academia: opportunities and challenges. *BioData Min* 16:1–11
- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. *Nat Med* 29:1930–1940
- Sakirin T, Ben Said R (2023) User preferences for ChatGPT-powered conversational interfaces versus traditional methods. *MJCSC*. <https://doi.org/10.58496/MJCSC/2023/004>
- 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* 6:1–5
- Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM (2023) Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 183:589–596. <https://doi.org/10.1001/jamainternmed.2023.1838>
- Lee TC, Staller K, Botoman V, Pathipati MP, Varma S, Kuo B (2023) ChatGPT answers common patient questions about colonoscopy. *Gastroenterology* 165:509–511.e7
- 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* 25:e47479. <https://doi.org/10.2196/47479>
- Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL (2015) Access to care and use of the internet to search for health information: results from the US national health interview survey. *J Med Internet Res* 17:e106. <https://doi.org/10.2196/jmir.4126>
- Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J (2023) Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 25:1–9
- Kamiński M, Łoniewski I, Misera A, Marlicz W (2019) Heartburn-related internet searches and trends of interest across six western countries: a four-year retrospective analysis using google ads keyword planner. *Int J Environ Res Public Health* 16:1–15. <https://doi.org/10.3390/ijerph16234591>
- Beck F, Richard JB, Nguyen-Thanh V, Montagni I, Parizot I, Renahy E (2014) Use of the internet as a health information resource among French young adults: results from a nationally representative survey. *J Med Internet Res* 16:1–13. <https://doi.org/10.2196/jmir.2934>
- Mikalef P, Kourouthanassis PE, Pateli AG (2017) Online information search behaviour of physicians. *Health Info Libr J* 34:58–73. <https://doi.org/10.1111/hir.12170>
- Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G (2023) Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 29:1
- El-Serag HB, Sweet S, Winchester CC, Dent J (2014) Update on the epidemiology of gastro-oesophageal reflux disease: a systematic review. *Gut* 63:871–880
- 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* 118:1–4
- Slater BJ, Dirks RC, McKinley SK, Ansari MT, Kohn GP, Thosani N, Qumseya B, Billmeier S, Daly S, Crawford C, Ehlers PA, 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* 35:4903–4917. <https://doi.org/10.1007/s00464-021-08625-5>
- Moore M (2016) Gastroesophageal reflux disease: a review of surgical decision making. *World J Gastrointest Surg* 8:77. <https://doi.org/10.4240/wjgs.v8.i1.77>
- Sachs GF, Ourshalimian S, Jensen AR, Kelley-Quon LI, Padilla BE, Shew SB, Lofberg KM, Smith CA, Roach JP, Pandya SR, Russell KW, Ignacio RC (2023) Machine learning to predict pediatric choledocholithiasis: a western pediatric surgery research consortium retrospective study. *Surgery* 174:934–939
- Marcinkevičs R, Wolfertetter PR, Klimiene U, Chin-Cheong K, Paschke A, Zerres J, Denzinger M, Niederberger D, Wellmann S, Ozkan E, Knorr C, Vogt JE (2024) Interpretable and intervenable ultrasonography-based machine learning models for pediatric

- appendicitis. *Med Image Anal* 91:103042. <https://doi.org/10.5281/zenodo.7>
20. Emile SH, Ghareeb W, Elfeki H, El Sorogy M, Fouad A, Elrefai M (2022) Development and validation of an artificial intelligence-based model to predict gastroesophageal reflux disease after sleeve gastrectomy. *Obes Surg* 32:2537–2547. <https://doi.org/10.1007/s11695-022-06112-x>
  21. Ge Z, Wang B, Chang J, Yu Z, Zhou Z, Zhang J, Duan Z (2023) Using deep learning and explainable artificial intelligence to assess the severity of gastroesophageal reflux disease according to the los angeles classification system. *Scand J Gastroenterol* 58:596–604. <https://doi.org/10.1080/00365521.2022.2163185>
  22. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A (2023) How AI responds to common lung cancer questions: ChatGPT vs google bard. *Radiology* 307:1–12. <https://doi.org/10.1148/radiol.230922>
  23. Bowman SR (2023) Eight things to know about large language models. *arXiv* 1–16.
  24. Eysenbach G (2023) The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ* 9:1–13
  25. Beaulieu-Jones BR, Shah S, Berrigan MT, Marwaha JS, Lai S-L, Brat GA (2024) Evaluating capabilities of large language models: performance of GPT4 on surgical knowledge assessments. *Surgery* 12:1–7. <https://doi.org/10.1016/j.surg.2023.12.014>
  26. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2:e0000198. <https://doi.org/10.1371/journal.pdig.0000198>
  27. Smith DA (2020) Situating wikipedia as a health information resource in various contexts: a scoping review. *PLoS ONE* 15:1–19. <https://doi.org/10.1371/journal.pone.0228786>
  28. Lee K, Hoti K, Hughes JD, Emmerton L (2014) Dr google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *J Med Internet Res* 16:1–14. <https://doi.org/10.2196/jmir.3706>
  29. Ayoub NF, Lee Y-J, Grimm D, Balakrishnan K (2023) Comparison between ChatGPT and google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg* 149:555–556
  30. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S (2023) ChatGPT vs google for queries related to dementia and other cognitive decline: comparison of results. *J Med Internet Res* 25:1–13. <https://doi.org/10.2196/48966>
  31. Mahajan A, Esper S, Oo TH, McKibben J, Garver M, Artman J, Klahre C, Ryan J, Sadhasivam S, Holder-Murray J, Marroquin OC (2023) Development and validation of a machine learning model to identify patients before surgery at high risk for postoperative adverse events. *JAMA Netw Open* 6:E2322285. <https://doi.org/10.1001/jamanetworkopen.2023.22285>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Bright Huo<sup>1</sup> · Elisa Calabrese<sup>2</sup> · Patricia Sylla<sup>3</sup> · Sunjay Kumar<sup>4</sup> · Romeo C. Ignacio<sup>5</sup> · Rodolfo Oviedo<sup>6,7,8</sup> · Imran Hassan<sup>9</sup> · Bethany J. Slater<sup>10</sup> · Andreas Kaiser<sup>11</sup> · Danielle S. Walsh<sup>12</sup> · Wesley Vosburg<sup>13</sup> 

✉ Wesley Vosburg  
wesvosburg@gmail.com

<sup>1</sup> Division of General Surgery, Department of Surgery, McMaster University, Hamilton, ON, Canada

<sup>2</sup> University of California South California, East Bay, Oakland, CA, USA

<sup>3</sup> Division of Colon and Rectal Surgery, Department of Surgery, Icahn School of Medicine at Mount Sinai, New York, NY, USA

<sup>4</sup> Department of General Surgery, Thomas Jefferson University Hospital, Philadelphia, PA, USA

<sup>5</sup> Division of Pediatric Surgery/Department of Surgery, San Diego School of Medicine, University of California, California, CA, USA

<sup>6</sup> Nacogdoches Center for Metabolic and Weight Loss Surgery, Nacogdoches, TX, USA

<sup>7</sup> University of Houston Tilman J. Fertitta Family College of Medicine, Houston, TX, USA

<sup>8</sup> Sam Houston State University College of Osteopathic Medicine, Conroe, TX, USA

<sup>9</sup> University of Iowa, Iowa City, IA, USA

<sup>10</sup> Department of Surgery, University of Chicago, Chicago, IL, USA

<sup>11</sup> Division of Colorectal Surgery, Department of Surgery, City of Hope National Medical Center, Duarte, CA, USA

<sup>12</sup> Department of Surgery, University of Kentucky, Lexington, KY, USA

<sup>13</sup> Department of Surgery, Harvard Medical School, Mount Auburn Hospital, Cambridge, MA, USA