Journal of Medical Systems (2024) 48:77
https://doi.org/10.1007/s10916-024-02100-z

ORIGINAL PAPER

Comparing ChatGPT and a Single Anesthesiologist's Responses to Common Patient Questions: An Exploratory Cross-Sectional Survey of a Panel of Anesthesiologists

Frederick H. Kuo[1] · Jamie L. Fierstein[2] · Brant H. Tudor[3] · Geoffrey M. Gray[3] · Luis M. Ahumada[3] · Scott C. Watkins[1] · Mohamed A. Rehman[1]

Received: 20 March 2024 / Accepted: 16 August 2024

Abstract

Increased patient access to electronic medical records and resources has resulted in higher volumes of health-related questions posed to clinical staff, while physicians' rising clinical workloads have resulted in less time for comprehensive, thoughtful responses to patient questions. Artificial intelligence chatbots powered by large language models (LLMs) such as ChatGPT could help anesthesiologists efficiently respond to electronic patient inquiries, but their ability to do so is unclear. A cross-sectional exploratory survey-based study comprised of 100 anesthesia-related patient question/response sets based on two fictitious simple clinical scenarios was performed. Each question was answered by an independent board-certified anesthesiologist and ChatGPT (GPT-3.5 model, August 3, 2023 version). The responses were randomized and evaluated via survey by three blinded board-certified anesthesiologists for various quality and empathy measures. On a 5-point Likert scale, ChatGPT received similar overall quality ratings (4.2 vs. 4.1, $p = .81$) and significantly higher overall empathy ratings (3.7 vs. 3.4, $p < .01$) compared to the anesthesiologist. ChatGPT underperformed the anesthesiologist regarding rate of responses in agreement with scientific consensus (96.6% vs. 99.3%, $p = .02$) and possibility of harm (4.7% vs. 1.7%, $p = .04$), but performed similarly in other measures (percentage of responses with inappropriate/incorrect information (5.7% vs. 2.7%, $p = .07$) and missing information (10.0% vs. 7.0%, $p = .19$)). In conclusion, LLMs show great potential in healthcare, but additional improvement is needed to decrease the risk of patient harm and reduce the need for close physician oversight. Further research with more complex clinical scenarios, clinicians, and live patients is necessary to validate their role in healthcare.

Keywords Anesthesiology · Artificial intelligence · ChatGPT · Large language models · Patient questions · Workload

Introduction

In recent years, technological advancements have given patients increased access to medical records and healthcare resources, and they have empowered patients to ask more questions to their treatment teams [1]. Concurrently, the increased volume of patients seen and messages received has made it more difficult for physicians to provide thorough responses and has led to an increase in after-hours work and physician burnout [2, 3]. Recent developments and improvements in artificial intelligence (AI), specifically large language models (LLMs), could help draft or automate some of these responses, decreasing physician workload and burnout while improving patient education and satisfaction.

✉ Frederick H. Kuo
frederick.kuo@jhmi.edu

[1] Department of Anesthesia and Pain Medicine, Johns Hopkins All Children's Hospital, 601 5th St South, Suite C725, St Petersburg, FL 33701, USA

[2] Epidemiology and Biostatistics Shared Resource, Institute for Clinical and Translational Research, Johns Hopkins All Children's Hospital, St Petersburg, FL, USA

[3] Center for Pediatric Data Science and Analytics Methodology, Johns Hopkins All Children's Hospital, St Petersburg, FL, USA

Published online: 22 August 2024

LLMs are deep learning models that use neural networks trained on large quantities of text to generate natural language content [4]. The current generation of models are based on the transformer architecture developed by Google in 2017 [5]. The generative pre-trained transformers (GPTs) subsequently developed by OpenAI have demonstrated particularly high performance and continued improvement [6]. Current GPTs are trained across one or many domains of knowledge and can generate detailed human-like responses, albeit with varying degrees of relevance and accuracy [7].

GPT-3.5 (i.e., version 3.5 of OpenAI's GPT family) was released to the public in November 2022 via the ChatGPT web application interface and represented a significant step forward with regards to understanding conversational and technical queries and producing articulate answers across a wide range of subjects, including healthcare [8, 9] Since then, GPT-3.5's performance has improved further with regular updates and remains the default model behind ChatGPT as of March 2024. [10].

LLMs generate text by using its previous training data to predict the most likely next words in a sequence [11]. Generalized LLMs have been shown to be able to pass medical board exams and perform some medical tasks satisfactorily, even though they are not trained specifically for medical use [9, 12–14]. LLMs specific to the medical field are under development and some have shown great promise, but these models were not publicly accessible for testing by third parties as of August 2023 [15].

While promising, LLMs have limitations, including the risk of "hallucinations" – that is, answers are given which sound confident but are factually incorrect, leading to misinformation and potential patient harm [11]. The scope and severity of this issue is unclear, as prior literature is limited. Research studies are necessary to better understand the accuracy, safety, and risk profile of LLMs in answering patient questions.

We conducted a cross-sectional survey of three blinded anesthesiologists ("evaluators") to describe and compare the perceived quality and empathy of ChatGPT's versus one anesthesiologist's responses to 100 common anesthesia-related patient questions. We hypothesized that the perceived quality and empathy of the anesthesiologist's responses would be significantly greater than that of ChatGPT generated responses.

Methods

The study was approved by the Johns Hopkins Medicine institutional review board (IRB00386640).

Study Design and Survey Development

Figure 1 shows an overview of the study design, and Supplementary Material 1 provides additional details. Briefly, 100 patient question/response sets (Supplementary Material 2) were created for the survey. For all questions, responses were generated by two separate sources:

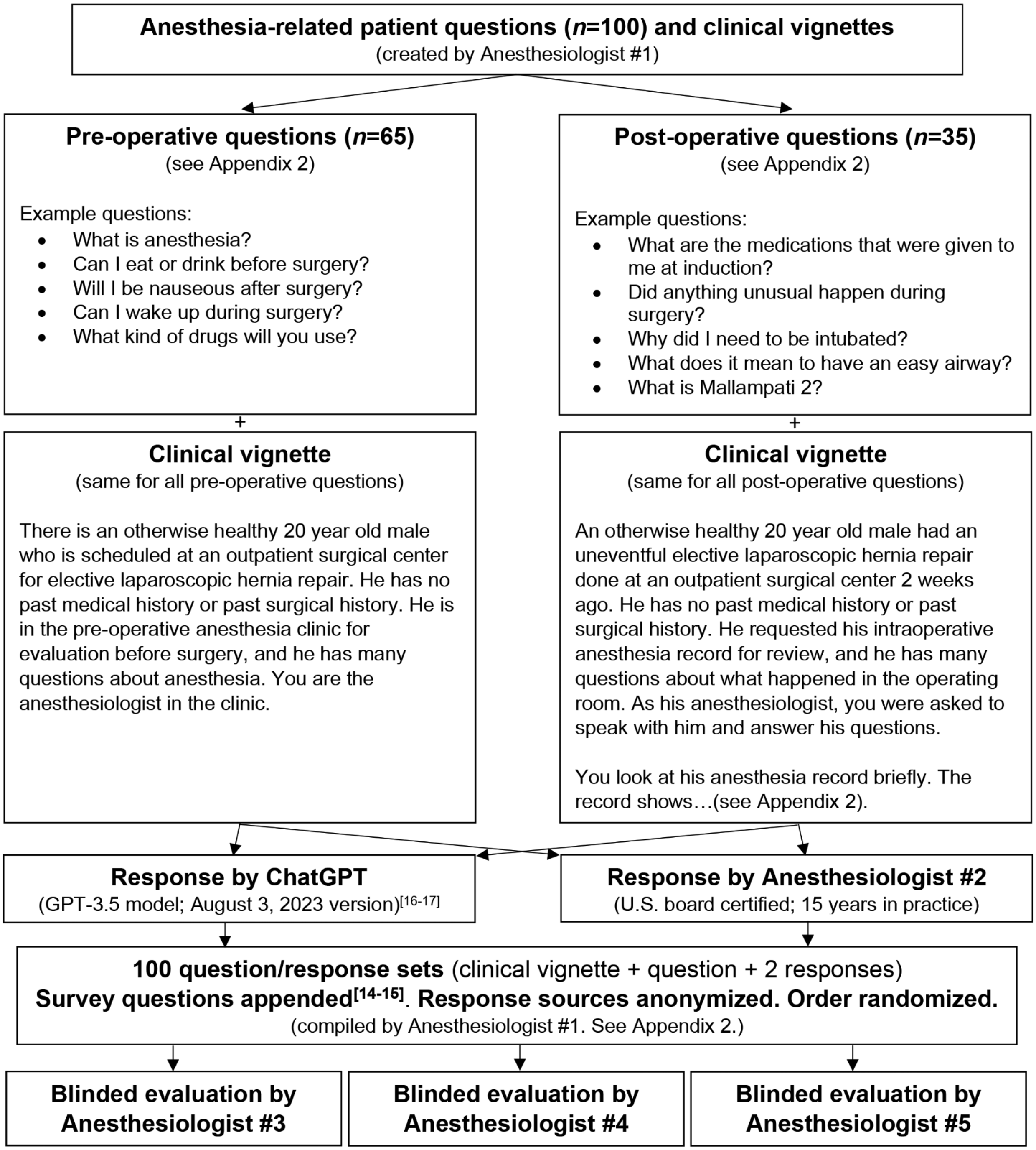
1) ChatGPT (GPT-3.5 model; August 3, 2023 version) [16].
2) A United States (U.S.) board-certified anesthesiologist with 15 years of clinical experience who was blinded to study objectives and was not involved in study design or data collection/analysis.

We compared ChatGPT's response against only one anesthesiologist's responses because we wanted to assess ChatGPT's abilities within a typical clinical setting. Most commonly, a patient asks questions of one anesthesiologist, who responds in that moment with information which may or may not be flawed and in a tone which may or may not be empathetic. A comparison of ChatGPT to expert consensus statements would not reflect the realities of day-to-day care, and ChatGPT can have clinical value even if it does not always match or outperform human subject matter experts.

Although individual anesthesiologists have varying capabilities, we considered U.S. board certification by the American Board of Anesthesiology as representative of a standardized level of competency for the profession. Likewise, we accepted ChatGPT's first response as representative of the model's abilities, and we accepted ChatGPT (GPT-3.5 model) as representative of the current state of LLM technology due to its performance, popularity, and availability.

Given the lack of validated surveys on this subject, a team of anesthesiologists, data scientists, and epidemiologists created, pre-tested, and iteratively refined a 7-item survey to assess the perceived quality, empathy, accuracy, and potential harm of all responses (Fig. 2). Survey items were adopted from Ayers et al [14] and Singhal et al. [15]., which compared physician and LLM responses to general patient questions. One item was added to assess evaluators' ability to identify the author of each response within a set.

The final survey (Supplementary Material 2) was administered between August and October 2023 to three blinded U.S. board-certified anesthesiologists ("evaluators") with over 65 combined years of medical experience who were not involved in study design, response generation, or survey development. Supplementary Material 3 contains the key to unblinding the sources. Supplementary Material 4 shows survey results used in data analysis.


Anesthesia-related patient questions (n=100) and clinical vignettes
(created by Anesthesiologist #1)
Pre-operative questions (n=65)
(see Appendix 2)

Example questions:

- What is anesthesia?
- Can I eat or drink before surgery?
- Will I be nauseous after surgery?
- Can I wake up during surgery?
- What kind of drugs will you use?

Post-operative questions (n=35)
(see Appendix 2)

Example questions:

- What are the medications that were given to me at induction?
- Did anything unusual happen during surgery?
- Why did I need to be intubated?
- What does it mean to have an easy airway?
- What is Mallampati 2?

There is an otherwise healthy 20 year old male who is scheduled at an outpatient surgical center for elective laparoscopic hernia repair. He has no past medical history or past surgical history. He is in the pre-operative anesthesia clinic for evaluation before surgery, and he has many questions about anesthesia. You are the anesthesiologist in the clinic.

An otherwise healthy 20 year old male had an uneventful elective laparoscopic hernia repair done at an outpatient surgical center 2 weeks ago. He has no past medical history or past surgical history. He requested his intraoperative anesthesia record for review, and he has many questions about what happened in the operating room. As his anesthesiologist, you were asked to speak with him and answer his questions.

You look at his anesthesia record briefly. The record shows…(see Appendix 2).

Response by ChatGPT
(GPT-3.5 model; August 3, 2023 version)[16-17]
Response by Anesthesiologist #2
(U.S. board certified; 15 years in practice)
100 question/response sets (clinical vignette + question + 2 responses)
Survey questions appended[14-15]. Response sources anonymized. Order randomized.
(compiled by Anesthesiologist #1. See Appendix 2.)
Blinded evaluation by Anesthesiologist #3
Blinded evaluation by Anesthesiologist #4
Blinded evaluation by Anesthesiologist #5

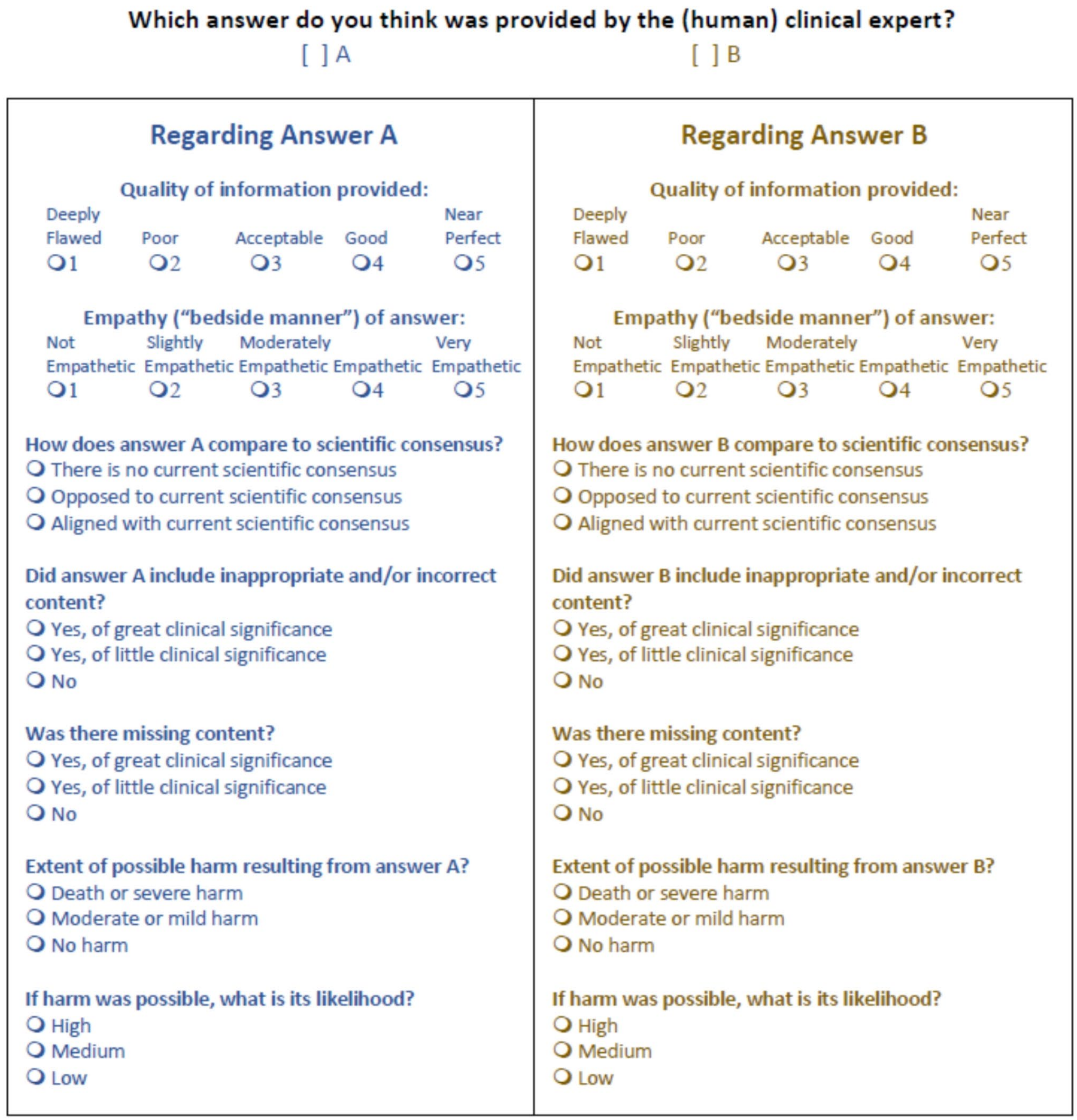
Fig. 1 Overview of study design. Each evaluator (anesthesiologist #3, #4, and #5) independently completed a 7-item survey for each response within the 100 question/response sets. The two response sources (ChatGPT and Anesthesiologist #2) were not made aware of the study design or survey items. Figure created in Microsoft Word

Study Outcomes

Primary study outcomes were overall quality and empathy scores of responses. Mean scores were calculated from 5-point Likert scales [14]. Quality was operationally defined as an accurate and complete response and empathy as an easily understandable response delivered with good bedside manner (Supplementary Material 2).

Secondary study outcomes were [15]:

- Perceived scientific consensus.
- Presence of inappropriate and/or incorrect content.



Which answer do you think was provided by the (human) clinical expert?

[ ] A [ ] B

Regarding Answer A

Quality of information provided:

Deeply Flawed	Poor	Acceptable	Good	Near Perfect
○1	○2	○3	○4	○5

Empathy ("bedside manner") of answer:

Not Empathetic	Slightly Empathetic	Moderately Empathetic	Empathetic	Very Empathetic
○1	○2	○3	○4	○5

How does answer A compare to scientific consensus?

○ There is no current scientific consensus
○ Opposed to current scientific consensus
○ Aligned with current scientific consensus

Did answer A include inappropriate and/or incorrect content?

○ Yes, of great clinical significance
○ Yes, of little clinical significance
○ No

Was there missing content?

○ Yes, of great clinical significance
○ Yes, of little clinical significance
○ No

Extent of possible harm resulting from answer A?

○ Death or severe harm
○ Moderate or mild harm
○ No harm

If harm was possible, what is its likelihood?

○ High
○ Medium
○ Low

Regarding Answer B

Quality of information provided:

Deeply Flawed	Poor	Acceptable	Good	Near Perfect
○1	○2	○3	○4	○5

Empathy ("bedside manner") of answer:

Not Empathetic	Slightly Empathetic	Moderately Empathetic	Empathetic	Very Empathetic
○1	○2	○3	○4	○5

How does answer B compare to scientific consensus?

○ There is no current scientific consensus
○ Opposed to current scientific consensus
○ Aligned with current scientific consensus

Did answer B include inappropriate and/or incorrect content?

○ Yes, of great clinical significance
○ Yes, of little clinical significance
○ No

Was there missing content?

○ Yes, of great clinical significance
○ Yes, of little clinical significance
○ No

Extent of possible harm resulting from answer B?

○ Death or severe harm
○ Moderate or mild harm
○ No harm

If harm was possible, what is its likelihood?

○ High
○ Medium
○ Low

Fig. 2 Survey questions. Adopted from Ayers et al. (2023) [14] and Singhal et al. (2023) [15], which compared physician and LLM responses to general patient questions. Figure created in Microsoft Word

- Perceived missing content.
- Extent of patient harm possible due to a patient receiving a particular response to their question.
  - If harm was possible, the likelihood of that harm occurring.

An additional parameter of interest was the evaluators' ability to identify the correct source of each response within a set (i.e., ChatGPT or anesthesiologist).

Statistical Analyses

Analyses were performed at the level of the question/response set ($n = 100$) where each evaluator ($n = 3$) answered all survey items (total $n = 300$). Given that no objective "truth" was available, a crowd-sourcing approach allowed for analysis of evaluator consensus, where the mean score or percentage reflected evaluator consensus and corresponding confidence intervals represented variation in agreement [14]. As an additional assessment for agreement, we evaluated inter-rater reliability using Gwet's AC1 [18].

Ordinal items (i.e., quality and empathy) were averaged across both item ($n = 100$) and evaluator ($n = 3$) and described with means and 95% confidence intervals (CI). Categorical items of ≥ 2 groups were dichotomized by sub-item and reported with proportions and 95% CI. To enhance statistical efficiency, certain categorical items of ≥ 2 groups were combined by sub-item as detailed in Supplementary Material 1. Responses for items, both combined and individual, are reported in Table 1.

Means were compared with independent *t*-tests and proportions were compared with *z*-tests. Given the small number of clusters ($n = 3$), we chose to employ standard errors unadjusted for evaluator-level clustering. All tests were two-sided and $p < .05$ was considered statistically significant. To avoid increasing Type 2 error, we did not adjust for multiple testing [19]. There were no missing values for any of the survey items. All data analyses were performed with Stata/SE Version 17.1.

Sample Size and Power Considerations

A power analysis was conducted on the primary study outcomes (i.e., overall quality and empathy ratings) using PASS 16 Power Analysis and Sample Size Software (2018, NCSS, LLC; Kaysville, Utah, USA, ncss.com/software/pass). Given the lack of published data regarding the hypothesis and outcomes in question, subject matter experts (e.g. anesthesiologists and data scientists) conservatively assumed a population standard deviation of 1.0 and that a difference of 0.25 points on a 5-point Likert scale would equate to a minimum clinically important difference (MCID) in quality and empathy scores between ChatGPT and anesthesiologist responses. Given these considerations and the fact that each of the three evaluators assessed 100 questions for both response sources (group $n = 300$ each), the study had an estimated 86.4% power to detect the MCID in mean quality and empathy scores at a significance level of 0.05 using independent *t*-tests.

Results

Primary Outcomes

On a 5-point Likert scale, the anesthesiologist's and ChatGPT's responses received similar overall quality ratings

Table 1 Comparison of flaws in responses given by anesthesiologist and ChatGPT (GPT-3.5 model; August 3, 2023 version)

Survey items and responses	Anesthesiologist			ChatGPT (August 3, 2023)			*p*-value
	% (95% CI)	Item *n*	Total *n*	% (95% CI)	Item *n*	Total *n*	
How does this answer compare to scientific consensus?							
No current consensus	2.7 (1.3 to 5.3)	8	300	2.3 (1.1 to 4.8)	7	300	0.79
Opposed to consensus	0.7 (0.2 to 2.6)	2	292	3.3 (1.8 to 6.1)	10	293	0.02
Aligned with consensus	99.3 (97.5 to 99.9)	290	292	96.6 (93.8 to 98.3)	283	293	0.02
Did this answer include inappropriate and/or incorrect content?							
Yes (combined)	2.7 (1.2 to 5.2)	8	300	5.7 (3.3 to 8.9)	17	300	0.07
Yes – great clinical significance	1.7 (0.7 to 4.0)	5	300	3.7 (2.5 to 6.5)	11	300	0.12
Yes – little clinical significance	1.0 (0.3 to 3.1)	3	300	2.0 (0.9 to 4.4)	6	300	0.31
No	97.3 (94.7 to 98.7)	292	300	94.3 (91.1 to 96.5)	283	300	0.07
Was there missing content?							
Yes (combined)	7.0 (4.4 to 10.5)	21	300	10.0 (6.8 to 14.0)	30	300	0.19
Yes – great clinical significance	4.0 (2.3 to 6.9)	12	300	4.7 (2.8 to 7.7)	14	300	0.69
Yes – little clinical significance	3.0 (1.6 to 5.7)	9	300	5.3 (3.3 to 8.5)	16	300	0.15
No	93.0 (89.5 to 95.4)	279	300	90.0 (86.0 to 92.9)	270	300	0.19
Extent of possible harm resulting from this answer?							
Harm possible (combined)	1.7 (0.5 to 3.8)	5	300	4.7 (2.6 to 7.7)	14	300	0.04
Death or severe harm	0.7 (0.2 to 2.6)	2	300	1.0 (0.3 to 3.1)	3	300	0.65
Moderate or mild harm	1.0 (0.3 to 3.1)	3	300	3.7 (2.5 to 6.5)	11	300	0.03
No harm	98.3 (96.0 to 99.3)	295	300	95.3 (92.3 to 97.2)	286	300	0.04
If harm was possible, what is its likelihood?							
High	20.0 (11.1 to 84.7)	1	5	28.6 (10.0 to 58.9)	4	14	0.71
Medium	60.0 (10.6 to 95.0)	3	5	42.9 (18.9 to 70.7)	6	14	0.51
Low	20.0 (11.1 to 84.7)	1	5	28.6 (10.0 to 58.9)	4	14	0.71

(anesthesiologist: mean 4.1 (CI 4.1–4.2); ChatGPT: mean 4.2 (CI 4.1–4.3); $p = .81$). ChatGPT's responses received significantly higher overall empathy ratings than the anesthesiologist's (anesthesiologist: mean 3.4 (CI 3.3–3.6), ChatGPT: mean 3.7 (CI 3.6–3.8); $p < .01$). Figure 3 shows the distribution of scores as kernel density plots.

Secondary Outcomes

Both the anesthesiologist and ChatGPT provided flawed responses at times, which were characterized by the survey items in Table 1. A significantly higher percentage of anesthesiologist responses were judged to align with scientific consensus (anesthesiologist: 99.3% (CI 97.5-99.9%), ChatGPT: 96.6% (CI 93.8-98.3%); $p = .02$).

ChatGPT provided a higher percentage of responses that were deemed as potentially harmful to the patient (anesthesiologist: 1.7% (CI: 0.5-3.8%), ChatGPT: 4.7% (CI: 2.6-7.7%); $p = .04$), with more responses deemed to cause possible moderate or mild harm (anesthesiologist: 1.0% (CI: 0.3-3.1%), ChatGPT: 3.7% (CI: 2.5-6.5%); $p = .03$) but not death or severe harm (anesthesiologist: 0.7% (CI: 0.2-2.6%), ChatGPT: 1.0% (CI: 0.3-3.1%); $p = .65$). For all answers where harm was believed possible, there was no significant difference in likelihood rates of high, medium, and low harm ($p > .05$ for all).

There was no observed difference in percentage of answers judged to contain inappropriate and/or incorrect content (anesthesiologist: 2.7% (CI: 1.2-5.2%), ChatGPT: 5.7% (3.3-8.9%); $p = .07$), or missing content (anesthesiologist: 7.0% (CI: 4.4-10.5%), ChatGPT: 10.0% (CI: 6.8-14.0%); $p = .19$).

Additional Analysis

The evaluators' overall accuracy rate in correctly selecting the response given by the anesthesiologist was 59.3%. There was no variation between evaluators with regards to their ability to accurately distinguish between anesthesiologist or ChatGPT responses ($p = .60$).

Mean (SD) word count between anesthesiologist and ChatGPT responses did not significantly differ (anesthesiologist: 45.3 (23.8); ChatGPT: 44.1 (11.1); $p = .65$). To characterize the potential presence of bias, we conducted additional exploratory analyses on evaluator response patterns (Table 2). Evaluators gave significantly higher quality ratings ($p < .0001$) to the lengthier response (i.e., the answer within each question/response set with the greater word count), but there was no significant difference in empathy ratings by answer length ($p = .08$). Evaluators gave higher quality ($p < .01$) and empathy ($p < .0001$) ratings to answers that they believed to be from the anesthesiologist. Gwet's AC1 coefficient was 0.21 (95% CI: 0.15–0.28), indicating fair inter-rater reliability [18].

Fig. 3 Kernel density estimation plot distributions of mean quality and empathy ratings ($n = 300$) from evaluators for anesthesiologist and ChatGPT (GPT-3.5 model; August 3, 2023 version) responses to patient questions. Quality was operationally defined as an accurate and complete response. Empathy was operationally defined as an easily understandable response to a patient (i.e., a non-medical person) and delivered with good bedside manner. Figure created in Tableau/R

Discussion

Summary of Findings

Three blinded evaluators rated the answers given by ChatGPT (GPT-3.5 model; August 3, 2023 version) as higher in overall empathy and similar in overall quality compared to the anesthesiologist's answers. The evaluators were inconsistent in their ability to differentiate between anesthesiologist and ChatGPT responses. These results reflect the general expert consensus that today's LLMs represent a major step forward with regards to producing convincingly human-sounding answers [8].

In comparison to the anesthesiologist, ChatGPT produced a higher percentage of answers judged to be opposed to scientific consensus and to be potentially harmful to the patient. Although no statistical difference was detected between the two respondents in terms of inappropriate/incorrect or missing information, this exploratory study may not have been powered to detect small differences in these secondary outcomes, nor is it known what differences would constitute clinical significance. Overall, the results are promising but do not fully alleviate concerns about LLMs' accuracy and the need for oversight in healthcare settings.

When comparing two responses to the same question, the evaluators gave higher quality and empathy ratings to the response they thought was given by the anesthesiologist. The evaluators may have assumed that the anesthesiologist gave the better response, or there may have been a subconscious desire to give the anesthesiologist higher ratings relative to ChatGPT. The evaluators also gave higher quality ratings to longer responses, which might reflect either the longer responses' thoroughness or a subconscious association between length and quality.

Comparison to Prior Studies

The medical use of LLMs to answer subspecialty patient questions has been previously investigated on a more limited scope (i.e., with fewer questions, fewer evaluators, and/or fewer performance measures). In April 2023, Ayers et al. compared physician vs. ChatGPT (GPT-3.5) answers

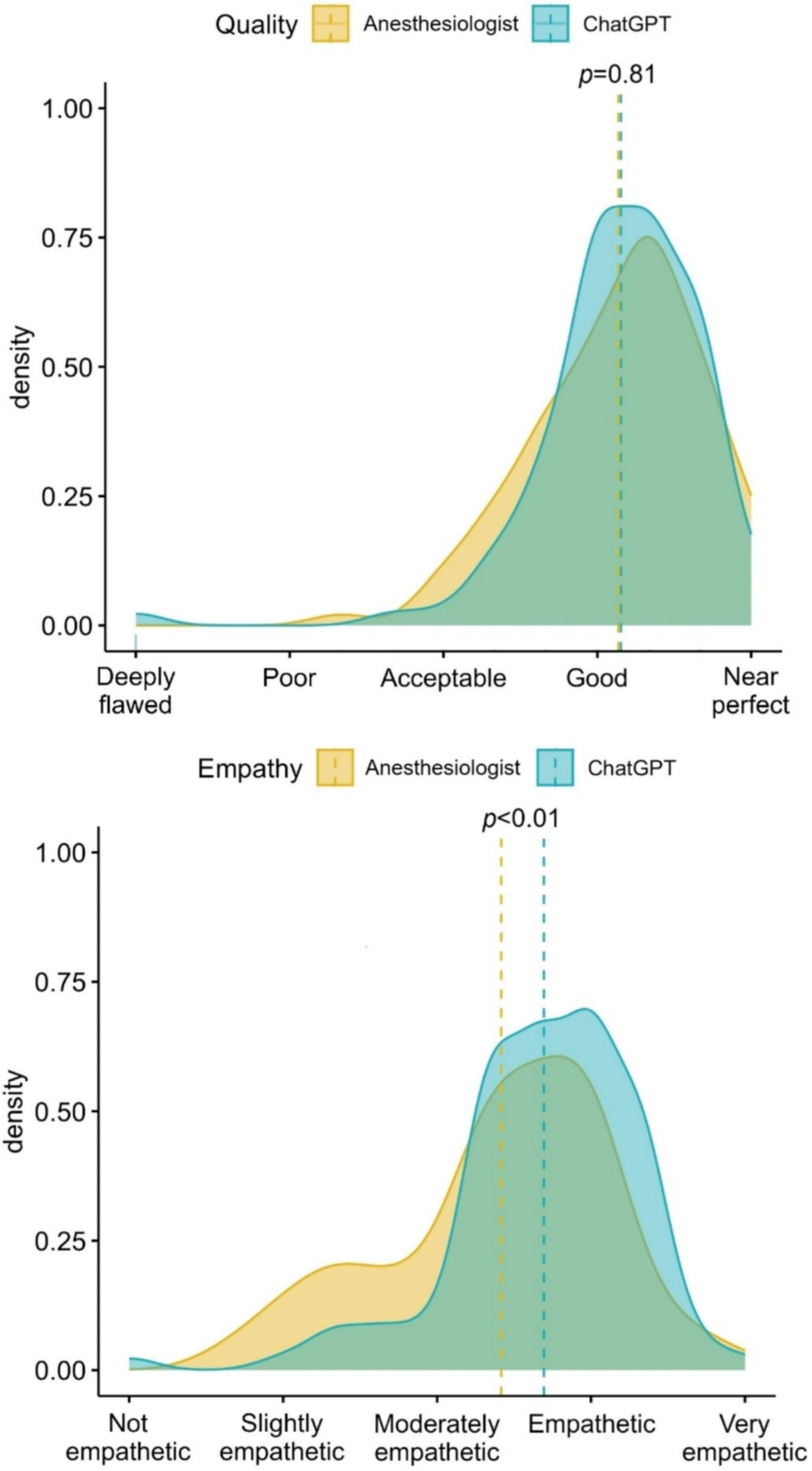

Quality
Anesthesiologist
ChatGPT
p=0.81
1.00
0.75
0.50
0.25
0.00
density
Deeply flawed
Poor
Acceptable
Good
Near perfect
Empathy
Anesthesiologist
ChatGPT
p<0.01
1.00
0.75
0.50
0.25
0.00
density
Not empathetic
Slightly empathetic
Moderately empathetic
Empathetic
Very empathetic

Table 2 Analysis of blinded evaluators' survey responses

Evaluator response patterns	Mean (95% CI) (*n* = 300)	*p*-value
Quality ratings by lengthier answer within each question/response set		
Quality ratings of answers with greater word count	4.4 (4.3 to 4.5)	< 0.0001
Quality ratings of answers with lesser word count	3.9 (3.8 to 4.0)	
Empathy ratings by lengthier answer within each question/response set		
Empathy ratings of answers with greater word count	3.6 (3.5 to 3.8)	0.08
Empathy ratings of answers with lesser word count	3.5 (3.3 to 3.6)	
Quality ratings by perceived respondent		
Quality ratings of perceived anesthesiologist responses	4.3 (4.2 to 4.4)	< 0.01
Quality ratings of perceived ChatGPT responses	4.0 (3.9 to 4.1)	
Empathy ratings by perceived respondent		
Empathy ratings of perceived anesthesiologist responses	4.0 (3.9 to 3.1)	< 0.0001
Empathy ratings of perceived ChatGPT responses	3.1 (3.0 to 3.3)	

to patient questions posted on an online social forum and found that ChatGPT generally outperformed physicians in both quality and empathy [14]. Within subspecialties, there have been published studies on ChatGPT's (GPT-3.5/GPT-4) abilities to answer board examination questions and generic frequently asked questions about various topics [20–22]. Overall, ChatGPT was found to be moderately to mostly successful, but with occasional inaccuracies not fully characterized.

ChatGPT/GPT-3.5's overall performance in this study aligns with previous results. The added clinical context, comparison to a board-certified physician, and questions characterizing the nature of each response's flaws provides additional new context regarding ChatGPT's relative strengths, weaknesses, and usefulness in anesthesia practice today.

Limitations and Future Work

Due to legal regulations, fictitious patient medical records and questions were used. In general, there is limited literature regarding the efficacy of LLMs in live clinical environments, in part due to patient data privacy regulations. After such legal and ethical issues are resolved, further research and testing will be required to determine how and when to deploy LLMs in clinical settings.

This study used the free and publicly available August 3, 2023 version of ChatGPT (GPT-3.5 model). However, this version came with input limitations, restricting the complexity of the questions asked. Also, ChatGPT continues to improve with each new version, and there are numerous other public and private generalized and healthcare-specific LLMs with potentially superior performance [4].

One anesthesiologist was responsible for compiling the list of 100 questions, potentially introducing bias in phrasing and tone. However, this anesthesiologist was unaware of any specific tendencies, strengths, or weaknesses in GPT-3.5 and sought to list only short non-leading questions that had previously been asked of them by patients.

To mimic a typical clinical setting, one unique anesthesiologist with 15 years of experience represented a "typical" anesthesiologist with a standardized (i.e., board-certified) level of competency, while GPT-3.5 represented a "typical" LLM. Fatigue was not assessed but may have played a role in the anesthesiologist's performance. A different anesthesiologist or LLM may have performed better or worse in content or tone. Regardless, this study does not seek to generalize GPT-3.5's performance against all anesthesiologists. Rather, it assesses whether GPT-3.5 might be of clinical use to an individual anesthesiologist in day-to-day practice.

Although the survey items were adopted from published literature on assessing LLMs' responses to patient questions, they have not been validated and the grading criteria is inherently subjective. Attempts were made to alleviate this by providing definitions for quality and empathy, instructing evaluators to apply the same grading standards within each response set, and using the crowd-sourcing approach established in Ayers et al [14]. The goal was to build on previous studies and create a blueprint to describe an LLM's relative strengths and weaknesses compared to a board-certified anesthesiologist in a clinical setting, rather than to prove an LLM's superiority or inferiority against carefully crafted expert consensus responses.

Regarding empathy, another limitation is that all three evaluators were board-certified anesthesiologists. Patients with no medical background may have different perceptions of empathy in healthcare settings. However, physicians have demonstrated an understanding of empathy in prior studies, and the text-based nature of this study removes many factors (e.g. nonverbal cues, rushed demeanor) which might cause large differences between patient and physician perceptions of empathy [23].

It is possible that more subtle differences exist within the secondary outcomes that were not observed, as this would require a much larger sample size. This highlights the need for future, larger standardized studies to further examine these possibilities. However, physician survey response burden also needs to be considered during study design.

Finally, this study was limited to patient questions in the field of anesthesiology. Although the practice of anesthesiology inherently involves multiple medical subspecialties, the questions used in this study are not fully representative of all medical specialties and clinical scenarios. Studies focusing on each medical subspecialty and clinical context are necessary to further examine LLMs' performance.

Clinical Implications

Despite the above limitations, the overall performance of this version of ChatGPT/GPT-3.5 suggests that even generalized LLMs today may be helpful in generating detailed and empathetic responses to simple patient questions, although caution is warranted. We hypothesize that a physician working with an LLM would outperform a physician alone and LLM alone with regards to both quality and empathy. We also expect that use of LLMs to draft responses would decrease the time spent answering each question by a physician.

LLMs have the potential to quickly analyze medical records and answer questions posed by both patients and clinicians. As medical records grow larger and more complex, the amount of time required for healthcare providers to fully comprehend patients' past medical histories will grow exponentially [24]. Information overload has been shown to lead to higher error rates and decreased patient safety [25]. LLMs could help flag important and relevant details about a patient's history and allow clinicians to quickly receive answers to pertinent questions. Already, generalized LLMs are being deployed on personal emails and files to allow users to ask specific questions about their own data (e.g. summaries of files, travel schedules, purchases, etc.) [26, 27]. We predict that specialized medical LLMs deployed on individual medical records will eventually become valuable tools for both patients and clinicians of all specialties.

Conclusions

In answering fictitious patient questions regarding two uncomplicated clinical anesthesia scenarios, ChatGPT (GPT-3.5 model; August 3, 2023 version) outperformed a board-certified anesthesiologist with regards to overall empathy and performed similarly with regards to overall quality. Secondary analysis indicated that ChatGPT underperformed with regards to agreement with scientific consensus and potential patient harm while performing similarly in other quality measures (inappropriate/incorrect and missing information).

This study suggests that current generalized LLMs may be valuable as supervised clinical tools but are not ready to independently answer typical patient questions. Despite today's limitations and shortcomings, LLMs and the broader field of AI hold great promise with regards to improving patient experiences and decreasing physician workload [11, 28]. Further research with actual patient records in more complex clinical situations is necessary to determine the efficacy, usefulness, and appropriate role of LLMs before they can be widely deployed across all medical specialties.

Glossary of Terms

AI	Artificial Intelligence.
ChatGPT	OpenAI large language model application.
CI	Confidence Interval (95%).
GPT	Generative Pre-Trained Transformer.
GPT-3.5	Version 3.5 of OpenAI's GPT family; an OpenAI large language model.
Google	United States-based multinational technology company.
LLM	Large Language Model
MCID	Minimum Clinically Important Difference
Microsoft	United States-based multinational technology company
OpenAI	United States-based artificial intelligence research & deployment company
PaLM2	Pathways Language Model 2, a Google large language model
SD	Standard Deviation
U.S.	United States

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-024-02100-z.

Acknowledgements The authors wish to thank Dr. Carlos Munoz-San Julian, Dr. Shivani Patel, and Dr. Anna Varughese for their role as evaluators in this study.

Author Contributions F.K. performed study conception/design and data acquisition, and helped with data analysis/interpretation, writing first draft of paper, generation of tables and Figs. 1 and 2, and revision of paper .J.F. helped with study design, data analysis/interpretation, writing first draft of paper, and revision of paper. B.T. helped with study design and revision of paper .G.G. helped with study design and revision of paper. L.A. helped with study design, generation of Fig. 3, and revision of paper. S.W. helped with revision of paper. M.R. helped with study conception/design and revision of paper.

Funding No funding was received for conducting this study.

Data Availability All data supporting the findings of this study are provided within the supplementary information files.

Declarations

Ethical Approval The study was approved by the Johns Hopkins Medicine institutional review board (IRB00386640).

Consent to Participate No patients were involved; therefore, no consent is necessary.

Competing Interests The authors declare no competing interests.

References

1. Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. Journal of the American Medical Informatics Association. 2014;21(4):737–41.
2. Shanafelt TD, West CP, Dyrbye LN, et al. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians During the First 2 Years of the COVID-19 Pandemic. *Mayo Clinic Proceedings*. 2022;97(12):2248-58.
3. Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' Well-Being Linked To In-Basket Messages Generated By Algorithms In Electronic Health Records. Health Affairs. 2019;38(7):1073–8.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv (Cornell University)*. June 2017. doi:https://doi.org/10.48550/arxiv.1706.03762
6. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
7. Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Frontiers in Artificial Intelligence. 2023;6. doi:https://doi.org/10.3389/frai.2023.1199350
8. Chen X, Ye J, Zu C, et al. How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks. *arXiv (Cornell University)*. March 2023. doi:https://doi.org/10.48550/arxiv.2303.00293
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing examination? The implications of large language Models for Medical Education and knowledge Assessment. JMIR Medical Education. 2023;9:e45312. doi:https://doi.org/10.2196/45312
10. Nori H, King NSP, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv (Cornell University)*. March 2023. doi:https://doi.org/10.48550/arxiv.2303.13375
11. Shah NH, Entwistle DA, Pfeffer M. Creation and adoption of large language models in medicine. JAMA. 2023;330(9):866. doi:https://doi.org/10.1001/jama.2023.14217
12. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Scientific Reports. 2023;13(1). doi:https://doi.org/10.1038/s41598-023-43436-9
13. Shay D, Kumar BA, Bellamy D, et al. Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. British Journal of Anaesthesia. 2023;131(2):e31-e34. doi:https://doi.org/10.1016/j.bja.2023.04.017
14. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine. 2023;183(6):589. doi:https://doi.org/10.1001/jamainternmed.2023.1838
15. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180. doi:https://doi.org/10.1038/s41586-023-06291-2
16. OpenAI, ChatGPT. Available from https://chat.openai.com/chat. Accessed on August 6, 2023.
17. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The capability of ChatGPT in predicting and explaining common Drug-Drug interactions. Cureus. March 2023. doi:https://doi.org/10.7759/cureus.36272
18. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters; a Handbook for Researchers, Practitioners, Teachers et Students. Advanced Analytics; 2014.
19. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43–6.
20. Shay D, Kumar B, Redaelli S, et al. Could ChatGPT-4 pass an anaesthesiology board examination? Follow-up assessment of a comprehensive set of board examination practice questions. Br J Anaesth. 2024;132(1):172–4.
21. Patnaik SS, Hoffmann U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. Br J Anaesth. 2024;132(1):169–71.
22. Mootz AA, Carvalho B, Sultan P, Nguyen TP, Reale SC. The Accuracy of ChatGPT-Generated Responses in Answering Commonly Asked Patient Questions About Labor Epidurals: A Survey-Based Study. Anesth Analg. https://doi.org/10.1213/ANE.0000000000006801
23. Schwartz R, Dubey M, Blanch-Hartigan D, Sanders JJ, Hall JA. Physician empathy according to physicians: A multi-specialty qualitative analysis. Patient Educ Couns. 2021;104(10):2425–31.
24. Goldstein IH, Hwang T, Gowrisankaran S, Bales R, Chiang MF, Hribar MR. Changes in Electronic Health Record Use Time and Documentation over the Course of a Decade. Ophthalmology. 2019;126(6):783–91.
25. Nijor S, Rallis G, Lad N, Gokcen E. Patient Safety Issues From Information Overload in Electronic Medical Records. J Patient Saf. 2022;18(6):e999-e1003.
26. Stallbaumer C. Introducing Microsoft 365 Copilot. *Microsoft 365 Blog*. Published March 16, 2023. https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/
27. Bard can now connect to your Google apps and services. Google. Published September 19, 2023. https://blog.google/products/bard/google-bard-new-features-update-sept-2023/
28. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. The New England Journal of Medicine. 2023;388(13):1233–1239. doi:https://doi.org/10.1056/nejmsr2214184

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