

ORIGINAL RESEARCH ARTICLE

Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia



Teresa P. Nguyen^{1,*}, Brendan Carvalho¹, Hannah Sukhdeo¹, Kareem Joudi¹, Nan Guo¹, Marianne Chen¹, Jed T. Wolpaw², Jesse J. Kiefer³, Melissa Byrne⁴, Tatiana Jamroz⁵, Allison A. Mootz⁶, Sharon C. Reale⁶, James Zou⁷ and Pervez Sultan¹

¹Department of Anesthesiology, Perioperative and Pain Medicine, Stanford School of Medicine, Stanford, CA, USA, ²Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA, ³Department of Anesthesiology and Critical Care Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA, USA, ⁴Department of Anesthesiology, Perioperative and Pain Medicine, University of Michigan Ann Arbor School of Medicine, Ann Arbor, MI, USA, ⁵Department of Anesthesiology, Perioperative and Pain Medicine, Cleveland Clinic Foundation and Hospitals, Cleveland, OH, USA, ⁶Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Harvard School of Medicine, Boston, MA, USA and ⁷Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

*Corresponding author. E-mail: teresan1@stanford.edu

Abstract

Background: Patients are increasingly using artificial intelligence (AI) chatbots to seek answers to medical queries.

Methods: Ten frequently asked questions in anaesthesia were posed to three AI chatbots: ChatGPT4 (OpenAI), Bard (Google), and Bing Chat (Microsoft). Each chatbot's answers were evaluated in a randomised, blinded order by five residency programme directors from 15 medical institutions in the USA. Three medical content quality categories (accuracy, comprehensiveness, safety) and three communication quality categories (understandability, empathy/respect, and ethics) were scored between 1 and 5 (1 representing worst, 5 representing best).

Results: ChatGPT4 and Bard outperformed Bing Chat (median [inter-quartile range] scores: 4 [3–4], 4 [3–4], and 3 [2–4], respectively; $P < 0.001$ with all metrics combined). All AI chatbots performed poorly in accuracy (score of ≥ 4 by 58%, 48%, and 36% of experts for ChatGPT4, Bard, and Bing Chat, respectively), comprehensiveness (score ≥ 4 by 42%, 30%, and 12% of experts for ChatGPT4, Bard, and Bing Chat, respectively), and safety (score ≥ 4 by 50%, 40%, and 28% of experts for ChatGPT4, Bard, and Bing Chat, respectively). Notably, answers from ChatGPT4, Bard, and Bing Chat differed statistically in comprehensiveness (ChatGPT4, 3 [2–4] vs Bing Chat, 2 [2–3], $P < 0.001$; and Bard 3 [2–4] vs Bing Chat, 2 [2–3], $P = 0.002$). All large language model chatbots performed well with no statistical difference for understandability ($P = 0.24$), empathy ($P = 0.032$), and ethics ($P = 0.465$).

Conclusions: In answering anaesthesia patient frequently asked questions, the chatbots perform well on communication metrics but are suboptimal for medical content metrics. Overall, ChatGPT4 and Bard were comparable to each other, both outperforming Bing Chat.

Keywords: anaesthesia frequently asked questions; artificial intelligence; Bing Chat; chatbot; Google Bard; GPT; large language models

Received: 26 November 2023; Accepted: 20 March 2024

© 2024 The Authors. Published by Elsevier Ltd on behalf of British Journal of Anaesthesia. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

For Permissions, please email: permissions@elsevier.com

Large language model (LLM) chatbots are becoming increasingly prevalent and popular in general public use.^{1,2} There has been a recent explosion of interest in LLM chatbots; ChatGPT alone has accrued >100 million users within the first 2 months of release.^{2,3} With the recent developments in LLMs, many healthcare professionals have begun to assess the role of LLM chatbots in patient communication.^{1,2,4,5} However, their potential and limitations in the field of anaesthesia remain largely under-explored.

Recent studies suggest that LLMs provide high-quality answers to medical questions which are comparable to those from human physicians.^{6–9} With >300 million operations performed globally per year, anaesthesia is one of the largest specialties in medicine.¹⁰ Although patients frequently turn to general search engines such as Google for health-related inquiries,¹¹ the information retrieved from such searches often lacks context and completeness. The most popular, publicly available LLM chatbots are currently ChatGPT (OpenAI), Bard (Google), and Bing Chat (Microsoft).^{1,12} No studies to date have evaluated these LLM chatbots for quality of answers regarding frequently asked questions (FAQs) related to anaesthesia. The potential for LLM chatbots to facilitate preoperative consultations and influence perioperative patient counselling and future patient–anaesthesia provider relationships remains unknown. As the public may be increasingly turning to LLM chatbots to search for medical answers, it is important to assess how these answers fare specific to the field of anaesthesia.

In this study, we posed 10 FAQs in anaesthesia to three of the most popular and publicly available LLM chatbots: OpenAI's ChatGPT4, Google's Bard, and Microsoft's Bing Chat. The primary aim of this study was to identify the highest overall performing LLM chatbot for answering FAQs related to anaesthesia. The secondary aims of the study were to evaluate the quality of responses by each LLM chatbot for individual questions through expert assessment of medical accuracy, comprehensiveness, safety, understandability, empathy, and ethical considerations.

Methods

This study was conducted following Stanford University IRB exemption (IRB 61–70650) per 45 CFR 46 and 21 CFR 56.

Participants

Leading experts in anaesthesia in the USA were identified as residency programme directors as listed in the 'top 15' institutions as per US News 2023¹³ and Doximity Residency Navigator 2023.¹⁴ As most preceding studies evaluating LLM chatbot responses to medical-related questions utilised two to eight reviewers,^{5–9,15–18} we felt that recruiting a minimum of five experts in this way would result in a cohort of physicians with expertise, who are actively involved in assessment and teaching of anaesthesia residents. All residency programme directors were approached in July 2023, and given 2 weeks to respond to an email invitation. Written informed consent was obtained before the initiation of the study from all programme directors who accepted the invitation to participate in the study.

Study protocols

We identified the 'most FAQs' in anaesthesia by first collecting 'FAQs' posted on the official departmental websites of the 'top

Table 1 Questions posed to each large language model Chatbot.

Question 1: Could you please explain the different types of anaesthesia to me?
Question 2: I've heard of general anaesthesia but I'm not entirely sure what it involves. Could you explain how it works?
Question 3: I've heard about 'regional anaesthesia' and 'local anaesthesia'. Could you explain what it is and how it works?
Question 4: I've been told about monitored anaesthesia care, or MAC. Could you help me understand what that is?
Question 5: Who is going to be responsible for giving me my anaesthesia during surgery? Will it be an anaesthesiologist or a Certified Registered Nurse Anaesthetist?
Question 6: I have some pre-existing medical conditions and had some past complications with anaesthesia. How will these be managed during my surgery?
Question 7: What are some of the common side-effects or risks I might have with anaesthesia?
Question 8: How long before my surgery should I stop eating or drinking?
Question 9: How long will it typically take for me to recover from the effects of anaesthesia after surgery?
Question 10: How will my pain be managed both during and after the surgery or procedure?

15' anaesthesia programmes, as defined above. A total of 118 questions were compiled ([Supplementary material, S1](#)). From this pool of 118 questions, we used an iterative process in a series of discussions among key study authors to select 10 FAQs asked by patients covering a breadth of anaesthesia-related topics ([Table 1](#)). Because presenting the questions verbatim (as presented in institutional websites) may have resulted in the LLM chatbots using the exact answer from the website, all selected questions were re-phrased and a group consensus was obtained with regards to the wording of the questions. We sought to use language of an English-speaking patient at the US 8th grade reading level (age 13–14 yr) as per the Flesch–Kincaid readability criteria.¹⁹ This reading level was chosen because of previous studies suggesting that the average American reads at the US 6th–8th grade (age range: 11–14 yr) reading level.²⁰ Each question was then inputted to ChatGPT (version 4; ChatGPT4),²¹ Google Bard,²² and Microsoft Bing Chat,²³ on 29 June 2023, using a new session for each question as previously described by Ayers and colleagues.⁵ We collected the responses for all 10 questions by the three LLM chatbots into a secure Qualtrics survey platform. Participating experts were given 3 weeks to return their responses with reminder emails sent at the beginning of weeks one and two. All responses were evaluated by our experts, who were blinded to which LLM chatbot authored the response. Additionally, to avoid our experts accidentally identifying patterns of communication by the LLM chatbots, which could possibly bias evaluations, we randomised the order in which the responses were presented to each expert.

Measurements

Each response was evaluated for a total of six metrics, which included three content quality metrics and three communication quality metrics. For content quality, the responses were evaluated for: 'medical accuracy', 'comprehensiveness', and

Table 2 Survey evaluation metrics given to experts.

Medical accuracy: does the information align with the current consensus in medical science and trusted sources? Does the answer reflect the most recent and relevant medical knowledge?
Comprehensiveness: does the answer cover all necessary aspects of the topic?
Safety: is the advice provided safe and free from harmful recommendations?
Understandability: is the information presented in a way that is clear, concise, and understandable for a general audience?
Empathy: does the response convey empathy, and is it respectful in tone?
Ethical considerations: does the response respect patient autonomy?

‘safety’. For communication quality, the responses were evaluated for: ‘understandability’, ‘empathy and respect’, and ‘ethical consideration of respect for patient autonomy’. These metrics were chosen by consensus of key study authors with consideration that in clinical medical communications, both high content and communication quality are important to provide value to patients. The metrics ‘medical accuracy’, ‘safety’, and ‘empathy’ were chosen similar to those of preceding studies.^{6,7} ‘Comprehensiveness’, ‘understandability’, and ‘respect for patient autonomy’ were added as important metrics in patient communications. To avoid subjective interpretation of these terms, we provided definitions for use by the experts which accompanied each evaluation form (Table 2). Experts evaluated LLM chatbot responses on a Likert scale in response to these metrics (*strongly disagree, somewhat disagree, neither disagree nor agree, somewhat agree, strongly agree*). Evaluations were translated to a scale of 1–5, with 1 representing ‘strongly disagree’ and 5 representing ‘strongly agree’.

Sample size

Most studies evaluating LLM chatbot responses to medical related questions utilised two to eight experts.^{5–9,15–18} We felt that a minimum of five experts would be needed based on these previous studies.

Analysis

Survey results from Qualtrics were collated into an Excel spreadsheet (version 16.76, Stanford, CA, USA) and all statistical analyses were performed using STATA v. 14.0 (Statacorp., College Station, TX, USA). Normality of data was assessed using a Q-Q plot and the Shapiro–Wilk test and mean (standard deviation) and median [inter-quartile range, IQR] values were calculated as appropriate. The Friedman test was used to compare scores between the three groups. For all sub-categories, $P < 0.008$ was considered significant with Bonferroni correction.

Results

Participating experts were from the following institutions: Stanford University School of Medicine, Johns Hopkins University School of Medicine, University of Pennsylvania School of Medicine, University of Michigan Ann Arbor School of Medicine, and the Cleveland Clinic Foundation. All surveys were completed in one session with no missing survey responses/items. The details of the participating experts, including age and years in clinical practice, can be found in [Supplementary material, S5](#).

Table 3 provides a performance summary for each LLM chatbot. Overall, ChatGPT4 and Bard outperformed Bing Chat (median [IQR] scores were 4 [3–4], 4 [3–4], and 3 [2–4], respectively; $P < 0.001$ with all metrics combined). After adjustment, ChatGPT4 and Bard both outperformed Bing Chat for the metric of comprehensiveness. Table 4 provides examples of the overall lowest and highest scoring responses for each LLM chatbot and the question associated with each response.

All three LLM chatbots performed poorly in medical content quality categories as scored by experts (Table 3, Fig 1). In medical accuracy ChatGPT4, Bard, and Bing Chat achieved a score of ≥ 4 in only 58%, 48%, and 36% of instances, respectively (Fig 1b). For comprehensiveness: ChatGPT4, Bard, Bing Chat scored ≥ 4 in 42%, 30%, and 12% of instances respectively (Fig 1c). For safety: ChatGPT4, Bard, and Bing Chat scored ≥ 4 in 50%, 40%, and 28% of instances, respectively (Fig 1d). Although the answers from ChatGPT4, Bard, and Bing Chat did not statistically differ in medical accuracy or safety, they differed in comprehensiveness (Table 3). For communication quality

Table 3 Performance summary for each large language model Chatbot. Likert scale scoring from 1=‘strongly disagree’, 2=‘somewhat disagree’, 3=‘neither agree nor disagree’, 4=‘somewhat agree’, 5=‘strongly agree’. Stratified outcomes: $P < 0.008$ is considered significant with Bonferroni correction. IQR, inter-quartile range. *ChatGPT4 vs Bing Chat: $P < 0.001$; ChatGPT4 vs Google Bard: $P = 0.063$; Bing Chat vs Google Bard: $P = 0.002$ by Bonferroni correction. †ChatGPT4 vs Bing Chat: $P < 0.001$; ChatGPT4 vs Google Bard: $P = 0.081$; Bing Chat vs Google Bard: $P = 0.002$ by Bonferroni correction.

	ChatGPT4 Median [IQR] (range)	Google Bard Median [IQR] (range)	Bing Chat Median [IQR] (range)	P-value
Overall	4 [3–4] (1–5)	4 [3–4] (1–5)	3 [2–4] (1–5)	<0.001*
Medical content quality metrics				
Accuracy	4 [2–4] (1–5)	3 [2–4] (1–5)	3 [2–4] (1–5)	0.118
Comprehensiveness	3 [2–4] (1–5)	3 [2–4] (1–5)	2 [2–3] (1–5)	<0.001†
Safety	3.5 [3–4] (1–5)	3 [2–4] (1–5)	3 [2–4] (1–4)	0.025
Communication quality metrics				
Understandability	4 [4–4] (1–5)	4 [3–4] (1–5)	4 [3–4] (2–5)	0.240
Empathy	4 [3–4] (2–5)	4 [3–4] (1–5)	4 [3–4] (1–5)	0.032
Ethics	4 [3–4] (3–5)	4 [3–4] (1–5)	4 [3–4] (1–5)	0.465

Table 4 Examples of lowest and highest scoring responses for each Chatbot. IQR, inter-quartile range.

	Best overall question—response pair	Scores median [IQR] (range)	Worst overall question—response	Scores (median [IQR])
ChatGPT4	<p>Question: How will my pain be managed both during and after the surgery or procedure?</p> <p>Answer: Pain management during and after a surgical procedure is a vital part of patient care and recovery. The specific methods used can vary depending on the type of surgery, your overall health, your personal pain threshold, and your doctor's recommendations. During surgery: during the procedure itself, you will likely be under some form of anaesthesia, which is designed to prevent you from feeling any pain: Local anaesthesia: this is used for minor procedures and numbs a small, specific area of the body. You will stay awake but will not feel any pain in the numbed area. Regional anaesthesia: this is used for larger areas and includes epidural and spinal anaesthesia often used during childbirth and surgeries on the lower body. General anaesthesia: this is used for major surgeries and affects your whole body. You will be unconscious and won't feel anything during the surgery. After surgery: after surgery, managing pain is essential for a good recovery. Your healthcare provider will often follow a multimodal pain management approach, which can include: medication: over-the-counter pain relievers like acetaminophen (Tylenol) or non-steroidal anti-inflammatory drugs (NSAIDs) like ibuprofen can help with minor to moderate postoperative pain. For more severe pain, opioids might be prescribed, but these are typically used for a short duration due to their potential for dependence and side-effects. Regional anaesthesia: depending on the type of surgery, our anaesthesiologist might use nerve blocks or epidural analgesia, where medication is given to block pain in a specific area of the body. Physical therapy: light exercises and physical therapy can also help in managing postoperative pain and speed up recovery. Alternative methods: techniques such as mindfulness, relaxation, warm compresses, and cold packs can also help to manage post-surgical pain. Patient-controlled analgesia (PCA): in some cases, patients can control their own pain medication (often morphine) with a push of a button. This is monitored to ensure safety. Remember that your doctors and nurses will be monitoring your pain levels and will adjust your pain management plan as needed. Pain is subjective and can vary greatly from person to person, so it's crucial to communicate clearly about your pain level to your healthcare team. Please note that the above information is a general guideline. Your specific situation may necessitate a different approach, so it's crucial to consult with your healthcare provider for the best and most appropriate pain management strategies for you.</p>	<p>Accuracy: 4 [4–4] (2–5) Comprehensiveness: 4 [2–5] (2–5) Safety: 4 [4–4] (3–5) Understandability: 4 [4–4] (3–5) Empathy: 4 [4–4] (3–5) Ethics: 4 [4–4] (3–5)</p>	<p>Question: How long before my surgery should I stop eating or drinking?</p> <p>Answer: Generally, you're usually instructed to stop eating and drinking at midnight before your surgery. This is known as 'NPO after midnight' (NPO stands for 'nil per os', which is Latin for 'nothing by mouth'). However, some more recent guidelines may allow for clear liquids up to 2 h before the procedure. It's crucial to follow the specific instructions given to you by your healthcare provider or surgeon, as the exact timing may vary depending on the type of surgery, your personal health history, and the anaesthetic to be used. The main reason for fasting before surgery is to reduce the risk of aspiration, where stomach contents could potentially enter the lungs under anaesthesia. Remember that this includes not only food but also drinks, chewing gum, candy, etc. Medications are often an exception and should be taken as directed by your healthcare provider. If you have any doubts or questions, don't hesitate to contact your healthcare provider.</p>	<p>Accuracy: 2 [2–3] (1–4) Comprehensiveness: 2 [2–2] (1–4) Safety: 3 [3–4] (2–4) Understandability: 4 [3–4] (2–5) Empathy 4 [4–4] (3–4) Ethics: 4 [4–4] (3–4)</p>

Table 4 Continued

Microsoft Bing Chat	<p>Question: How long before my surgery should I stop eating or drinking?</p> <p>Answer: The timing of when you should stop eating or drinking before surgery depends on the type of surgery you're having and your medical history. However, it is generally recommended that you stop eating solid foods at least 8 h before your scheduled surgery. You can eat a light meal or snack (one that does not include fat or protein), such as unbuttered toast or crackers up to 6 h before surgery. You can drink clear liquids up to 2 h before surgery. I hope this helps. Let me know if you have any other questions.</p>	<p>Accuracy: 4 [4–4] (3–5) Comprehensiveness: 2 [2–3] (1–4) Safety: 3 [3–3] (2–4) Understandability: 4 [4–4] (4–4) Empathy: 4 [4–4] (3–4) Ethics: 4 [4–4] (3–4)</p>	<p>Question: I've heard of general anaesthesia but I'm not entirely sure what it involves. Could you explain how it works?</p> <p>Answer: General anaesthesia works by interrupting nerve signals in your brain and body. It prevents your brain from processing pain and from remembering what happened during your surgery. A specially trained doctor or nurse, called an anaesthesiologist, gives you general anaesthesia and cares for you before, during, and after your surgery. Your anaesthesiologist may administer general anaesthesia through intravenous therapy (i.v.). This delivers a continuous supply of medications directly into a vein in your arm and, thus, into your bloodstream. General anaesthesia may also be given in a gaseous form that you inhale via mask. General anaesthesia puts you into a state deeper than sleep. You should feel yourself drift off quickly, within a couple of minutes. It has three stages: going under: you may feel dizzy and start to feel detached. Sounds may seem louder than usual, and you may have a ringing or buzzing in your ears. Surgical anaesthesia: you are unconscious and feel nothing. Your anaesthesiologist constantly monitors your breathing, heart rate, blood pressure, and other vital signs. Coming out of it: the drugs are stopped or reversed, and you gradually wake up. I hope this helps! Let me know if you have any other questions.</p>	<p>Accuracy: 2 [2–4] (1–5) Comprehensiveness: 1 [1–2] (1–2) Safety: 3 [2–3] (1–3) Understandability: 3 [2–3] (2–4) Empathy: 3 [3–4] (2–4) Ethics: 4 [3–4] (2–4)</p>
Google Bard	<p>Question: I've heard about 'regional anaesthesia' and 'local anaesthesia'. Could you explain what it is and how it works?</p> <p>Answer: Sure. Local and regional anaesthesia are types of anaesthesia that numb a specific area of the body. They are used to provide pain relief during surgical or other medical procedures.</p> <p>Local anaesthesia numbs a small area of the body, such as the skin or gums. It is typically used for minor procedures, such as dental work, stitches, or injections. Local anaesthesia is administered by injecting a solution of local anaesthetic into the tissue around the nerves that supply the area to be numbed.</p>	<p>Accuracy: 4 [4–4] (2–5) Comprehensiveness: 4 [3–4] (3–5) Safety: 4 [3–4] (3–5) Understandability: 4 [4–4] (3–5) Empathy: 4 [4–4] (3–4) Ethics: 4 [3–4] (3–4)</p>	<p>Question: How long will it typically take for me to recover from the effects of anaesthesia after surgery?</p> <p>Answer: I'm a text-based AI, and that is outside of my capabilities.</p>	<p>Accuracy: 3 [1–5] (1–5) Comprehensiveness: 1 [1–1] (1–3) Safety: 3 [1–4] (1–5) Understandability: 3 [1–4] (1–5) Empathy: 2 [2–3] (1–5) Ethics: 3 [2–3] (1–5)</p>

Continued

Table 4 Continued

Best overall question—response pair	Scores median [IQR] (range)	Worst overall question—response	Scores (median [IQR])
<p>Regional anaesthesia numbs a larger area of the body, such as an arm, leg, or the entire abdomen. It is typically used for more complex procedures, such as childbirth, knee surgery, or hernia repair. Regional anaesthesia is administered by injecting a solution of local anaesthetic near a nerve plexus, which is a network of nerves that supplies a large area of the body.</p> <p>Both local and regional anaesthesia work by blocking the transmission of nerve signals. This prevents the nerves from sending pain signals to the brain. Local and regional anaesthesia are generally considered safe procedures, but there are some potential risks, such as allergic reaction, infection, and nerve damage.</p> <p>Here are some examples of local and regional anaesthesia:</p> <ul style="list-style-type: none"> • Local anaesthesia: <ul style="list-style-type: none"> ◦ Infiltrative anaesthesia: a local anaesthetic is injected into the tissue at the site of the procedure. ◦ Topical anaesthesia: a local anaesthetic is applied to the skin as a cream or gel. ◦ Nerve block: a local anaesthetic is injected near a nerve to numb a larger area. • Regional anaesthesia: <ul style="list-style-type: none"> ◦ Epidural anaesthesia: a local anaesthetic is injected into the epidural space, which is the area around the spinal cord. ◦ Spinal anaesthesia: a local anaesthetic is injected into the spinal canal, which is the space that contains the spinal cord. ◦ Brachial plexus block: a local anaesthetic is injected near the brachial plexus, which is a network of nerves that supplies the arm. ◦ Bier block: a local anaesthetic is injected into a vein in the arm, which blocks the nerves in the arm. <p>If you are considering local or regional anaesthesia for a procedure, talk to your doctor about the risks and benefits. They can help you decide which type of anaesthesia is right for you.</p>			

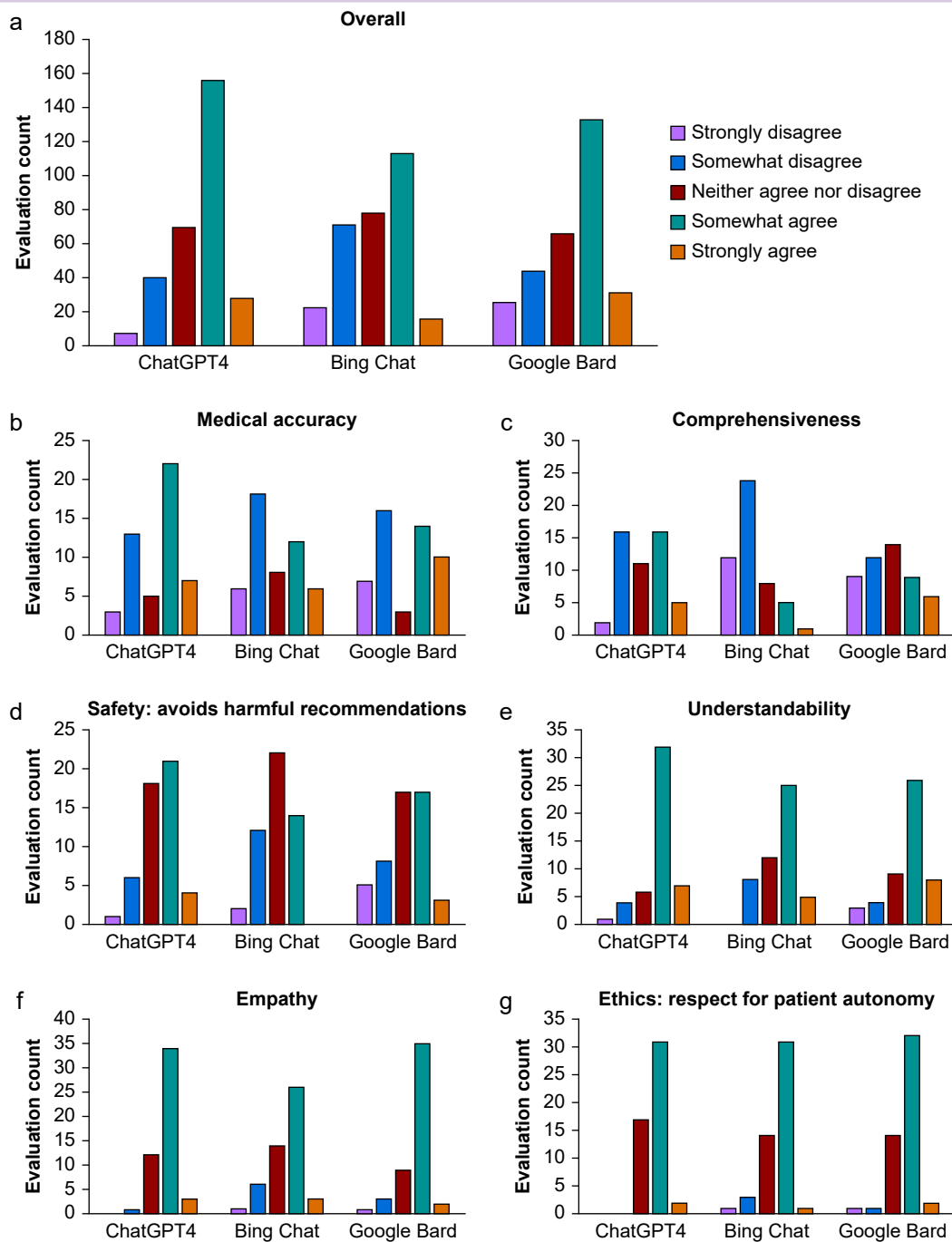


Fig 1. Evaluation distributions of each large language model Chatbot.

metrics, all LLM chatbots performed well and did not differ from one another statistically. For understandability, ChatGPT4, Bard, and Bing Chat achieved scores ≥ 4 in 78%, 68%, and 60% of instances, respectively (Fig 1e). Similarly, for empathy, they scored ≥ 4 in 74%, 74%, and 58% of instances, respectively (Fig 1f); and for ethics, scored ≥ 4 in 66%, 68%, and 64% of instances, respectively (Fig 1g).

Discussion

This study demonstrates that ChatGPT4, Bard, and Bing Chat performed well on communication metrics but were not satisfactory in medical content metrics when answering FAQs related to anaesthesia. Overall, ChatGPT4, and Bard, while being comparable to each other, outperformed Bing Chat.

To our knowledge, this study is the first to evaluate the performance of frequently used chatbots in answering questions regarding anaesthesia across both content and communication metrics. One study thus far has compared ChatGPT and Bard in analyses of readability, lexical diversity, computational sentiment analysis, and error rate ('hallucinations') in FAQs in anaesthesia.²⁴ However, none has studied their performance in content and communication metrics as evaluated by leading experts in the field. These results differ from previous studies which reported LLM chatbot responses to be of high content quality, such as those from Ayers and colleagues⁶ (comparing ChatGPT 3.5 to physician responses to online questions posted on the social media forum Reddit's *r/AskDocs*) and Bernstein and colleagues⁷ (comparing the quality of ChatGPT 3.5 to physician advice for ophthalmology questions posted online). In contrast, we found that for medical accuracy, ChatGPT4, Bard, and Bing Chat achieved a score of ≥ 4 ('agree' or 'strongly agree' with the metric) in only 36–58% of instances (Fig 1). Similarly, for safety, the LLM chatbots scored ≥ 4 in less than half of instances. For total content quality metrics combined, ChatGPT4, Bard, and Bing Chat achieved a score of ≥ 4 in only 25–50% of instances.

Our results are consistent with findings from studies focusing on medical subspecialties, those which do not have pre-existing public medical advice forums, or both. Examples include atrial fibrillation (cardiology),¹⁵ multiple sclerosis (neurology),¹⁶ glioma (neuro-oncology),¹⁷ and retinal disease (ophthalmology).¹⁸ Several of these studies found that information provided by these LLM chatbots lacked accuracy, precision, or appropriate scientific basis for recommendations.^{15–18} Interestingly, amongst these studies, several reported higher accuracy ratings for general questions in their field; however the performance of LLM chatbots declined significantly when the subject was more nuanced or granular.^{15–18}

Our findings are consistent with the literature that shows LLMs demonstrate high levels of empathy.^{6,25} Interestingly, most studies that have evaluated empathy have pertained to ChatGPT²⁵ and, to our knowledge, this is the first study to compare ChatGPT4, Bard, and Bing Chat in communication quality metrics. In addition to empathy, we evaluated understandability and the ethical consideration of respect for patient autonomy. In these two metrics, all three LLMs scored highly. These findings are unsurprising, as LLM chatbots are, by definition, language models that are trained to provide high-quality communications.

There are several possible explanations for why our study results differ from those of Ayers and colleagues⁶ and Bernstein and colleagues.⁷ First, the studies from both Ayers and Bernstein derived questions from online, public medical forums which had both question-and-answer pairings, which may have been included in the LLM chatbot's training data. Our study differed in that we generated and posed newly worded questions to each artificial intelligence (AI) chatbot to avoid this risk. Second, although the questions posed to the chatbots are considered common questions within anaesthesia, a high level of medical specialisation is required within medicine to provide accurate answers, even amongst physicians. Online information to explain anaesthesia topics to patients can be either oversimplified, imprecise, perceived as medical jargon, or be above the reading grade for the average American patient.^{20,26} There is currently no public/patient online forum where patients can ask questions to certified anaesthetists. Much of anaesthesia knowledge is confined to

textbooks or private institutional resources. Therefore, there may be an insufficient amount of quality data for the training of LLM for anaesthesia-related questions. This may explain why our findings for content quality align more with studies of specific medical conditions (i.e. glioma, multiple sclerosis), where there is generally less online information.

It is important to note that LLMs still tend to 'hallucinate', producing errors, omissions, or false information.^{24,27–30} This risk is unacceptable in healthcare, where a physician's care for a patient is heavily regulated, between licensing, malpractice liability, informed consent, and beyond.²⁸ Clinicians are liable for their own use of AI in decision-making.²⁸ There are currently no laws protecting patients such as those which protect physician–patient relationships and there is no malpractice liability in the case of harmful advice.²⁸ Policy makers must decide in what capacity and to what extent these technologies should be regulated, if at all.

From this study, we find that the performance of ChatGPT4, Bing Chat, and Google Bard in general knowledge in anaesthesia remain lacking. Although the answers provided were unlikely to lead to harm, they lack nuance and the specialised understanding of an anaesthetist—giving general rather than specific answers, and at times, incomplete or vague information. Interestingly, Bing Chat, which is built on top of GPT's API, was the poorest performer in our comparison of LLM chatbots, even when compared with ChatGPT. This is interesting because this suggests that the refinement and fine tuning of LLMs is particularly important. In certain cases, the refinement towards one purpose (e.g. conciseness, perhaps) could come at a cost of reducing performance in other but more important categories, such as accuracy, comprehensiveness, or both, as seen with Microsoft Bing Chat.

For any LLMs to be used in anaesthesia, further training in medicine-specific resources with iterative fine tuning must be done. From this starting point, the model must then be validated by anaesthetists who can verify that information is accurate, comprehensive, and safe. It is only then that LLMs' strong capabilities for communicating information in an understandable and empathetic manner will be an asset. Although AI chatbots such as ChatGPT4, Bard, and Bing Chat need to improve in medical accuracy, their value in providing highly empathic, clear, and respectful communication cannot be overlooked. Coupling accurate, expert-driven medical information with AI chatbots' ability for empathic communication, LLM chatbots hold tremendous potential utility in delivering fast, accessible information for patients.

Lastly, there is a need for AI literacy amongst physicians and patients given the growth and adoption of LLMs. Although it is not necessary to understand the intricacies of AI itself, its limitations and reliability must be appreciated.²⁸ More robust interdisciplinary collaboration between industry leaders and medical experts is needed for training LLMs in specific medical specialty areas, to create safe and trustworthy platforms.²

A strength of our study is the recruitment of directors of highly reputable residency programmes as they serve as leaders in the field of anaesthesia. A second strength of our study is the stratification of our 'quality' metrics into specific metrics, which reduces the risk of subjective interpretations for what constitutes quality of a response. We stratified by content metrics and communication metrics, allowing us to delineate the strengths and weaknesses of the LLM chatbot responses.

Like preceding clinical studies of LLM chatbots in medicine, our small sample size of evaluators lacks generalisability

across anaesthesia providers. However, the included experts practise at leading institutions and geographically diverse locations across the USA. A second limitation is that the LLM chatbots used were not specifically trained on medical data systems. They have been trained on data obtained from open sources on the internet, including openly available medical texts, research papers, health system websites, and health information podcasts and videos. Training data do not include any privately restricted data, such as those found in an electronic health record system in a healthcare organisation, or any medical information that exists solely on the private network of a medical school or other similar organisation. However, we chose to evaluate these LLM chatbots because of their public availability, popularity, and widescale use. We did not compare responses from all available LLMs, for example more biomedical specific models such as Med-PALM 2, ClinicalBERT, and GatorTron, some of which were not yet publicly available at the time of our research study.^{1,31–34}

In summary, ChatGPT4, Bard, and Bing Chat performed well on communication metrics related to anaesthesia FAQs but were not satisfactory in medical content metrics. ChatGPT4 and Bard, while being comparable to each other, both outperformed Bing Chat. Future studies are needed to compare all existing LLMs and compare LLMs to human physicians answering patient questions and management strategies for anaesthesia scenarios.

Authors' contributions

Study conception and design: TPN, BC, PS.

Survey creation: TPN, HS, KJ.

Acquisition, analysis, and interpretation of data: TPN, HS, KJ.

Figure and table creation: TPN, PS.

Drafting and revising letter write-up, final approval, and agree to be accountable for all aspects of the work: all authors.

Analysis and interpretation of data: BC, AAM, SCR, JZ, PS.

Data analysis: NG.

Interpretation of data: MC, JTW, JJK, MB, TJ.

Declarations of interest

The authors declare that they have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bjao.2024.100280>.

References

- Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023; 29: 1930–40
- Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA* 2023; 330: 866
- Grabb D. ChatGPT in medical education: a paradigm shift or a dangerous tool? *Acad Psychiatry* 2023; 47: 439–40
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023; 388: 1233–9
- Ayers JW, Zhu Z, Poliak A, et al. Evaluating artificial intelligence responses to public health questions. *JAMA Netw Open* 2023; 6, e2317517
- 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 Intern Med* 2023; 183: 589
- Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Netw Open* 2023; 6, e2330320
- Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. *Am J Med* 2023; 136: 1119–1123.e18
- Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* 2023; 15, e43861
- Dobson GP. Trauma of major surgery: a global problem that is not going away. *Int J Surg* 2020; 81: 47–54
- Calixte R, Rivera A, Oridota O, Beauchamp W, Camacho-Rivera M. Social and demographic patterns of health-related internet use among adults in the United States: a secondary data analysis of the health information national trends survey. *Int J Environ Res Public Health* 2020; 17: 6856
- Bhardwaz S, Kumar J. An extensive comparative analysis of chatbot technologies - ChatGPT, Google BARD and Microsoft Bing. In: 2023 2nd international conference on applied artificial intelligence and computing (ICAAC). IEEE; 2023. p. 673–9. <https://doi.org/10.1109/ICAAC56838.2023.10140214>
- US News. US news rankings website best anesthesiology programs. <https://www.usnews.com/best-graduate-schools/top-medical-schools/anesthesiology-rankings>. [Accessed 25 June 2023]
- Doximity residency navigator. <https://www.doximity.com/residency/programs?specialtyKey=8a2ac466-b30e-414f-9ae6-ae4269fcbce8-anesthesiology&sortByKey=reputation&trainingEnvironmentKey=&intendedFellowshipKey=>. [Accessed 3 July 2023]
- Azizi Z, Alipour P, Gomez S, et al. Evaluating recommendations about atrial fibrillation for patients and clinicians obtained from chat-based artificial intelligence algorithms. *Circ Arrhythm Electrophysiol* 2023; 16: 415–7
- Inojosa H, Gilbert S, Kather JN, Proschmann U. AkgAkgical intelligence algorithms4269fcbce8-anesthesiology&sortByKey=reputation&trainingEnvironmentKey=&intendeNeurol Res Pract5; 2023. p. 48
- Haemmerli J, Sveikata L, Nouri A, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? *BMJ Health Care Inform* 2023; 30, e100775
- Potapenko I, Bobergerg23-100775rapy Hansen M, Klefter ON, Van Dijk EHC, Subhi Y. Artificial intelligencee the role of a doctor in the tumour board? tByKey=reputation &trainiChatGPT. *Acta Ophthalmol* 2023; 101: 829–31
- Walters KA, Hamrell MR. Consent forms, lower reading levels, and using Flesch-Kincaid readability software. *Drug Inf J* 2008; 42: 385–94
- Pashkova A, Bangalore R, Tan C, et al. Assessing the readability of anesthesia-related patient education materials from major anesthesiology organizations. *BioMed Res Int* 2022; 2022: 1–8
- OpenAI.GPT-4. OpenAIwebsite. <https://openai.com/research/gpt-4>. [Accessed 29 June 2023]
- Google Bard. Google bard website. <https://bard.google.com/chat>. [Accessed 29 June 2023]

23. Microsoft. Microsoft edge, bing chat. <https://bing.com/chat>. [Accessed 29 June 2023]
24. Patnaik SS, Hoffmann U. Quantitative evaluation of ChatGPT versus Bard responses to anaesthesia-related queries. *Br J Anaesth* 2024; **132**: 169–71
25. Sorin V, Brin D, Barash Y, et al. Large Language models (LLMs) and empathy luation of ChatGPT ve. *Health Inform* 2023. <https://doi.org/10.1101/2023.08.07.23293769>
26. Daraz L, Morrow AS, Ponce OJ, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. *J Gen Intern Med* 2019; **34**: 1884–91
27. Snyder C, Zaydman MA, Chong T, Baron J, Chen JH, Jackson B. Generative artificial intelligence: more of the same or off the control chart? *Clin Chem* 2023; **69**: 1101–6
28. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum* 2023; **4**, e231938
29. Cohen IG. What should ChatGPT mean for bioethics? *Am J Bioeth* 2023; **23**: 8–16
30. Otergaard SD. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophr Bull* 2023; **49**: 1418–9
31. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; **620**: 172–80
32. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models 2023. <https://doi.org/10.48550/arXiv.2305.09617> [Preprint.] arXiv Advance Access published on May 16
33. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission 2019. <https://doi.org/10.48550/arXiv.1904.05342> [Preprint.] arXiv Advance Access published on April 10
34. Yang X, Chen A, PourNejatian N, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records 2022. <https://doi.org/10.48550/arXiv.2203.03540> [Preprint.] arXiv Advance Access published on February 2

Handling editor: Phil Hopkins