Comparison of ChatGPT vs. Bard to Anesthesia-related Queries

Sourav S. Patnaik, Ulrike Hoffmann

Affiliations:

Department of Anesthesiology and Pain Management, The University of Texas Southwestern Medical Center

Address Correspondence to:

Sourav S. Patnaik, PhD
Department of Anesthesiology and Pain Management
The University of Texas Southwestern Medical Center
5323 Harry Hines Blvd.
Dallas, TX 75390

Running Title: ChatGPT vs. Bard from anesthesiology perspective

Keywords: Anesthesia, patient, communication, ChatGPT, Bard, analytics

Abstract

We investigated the ability of large language models (LLMs) to answer anesthesia related queries prior to surgery from a patient's point of view. In the study, we introduced textual data evaluation metrics, investigated "hallucinations" phenomenon, and evaluated feasibility of using LLMs at the patient-clinician interface. ChatGPT was found to be lengthier, intellectual, and effective in its response as compared to Bard. Upon clinical evaluation, no "hallucination" errors were reported from ChatGPT, whereas we observed a 30.3% error in response from Bard. ChatGPT responses were difficult to read (college level difficulty) while Bard responses were more conversational and about 8th grade level from readability calculations. Linguistic quality of ChatGPT was found to be 19.7% greater for Bard (66.16 ± 13.42 vs. 55.27 ± 11.76; p=0.0037) and was independent of response length. Computational sentiment analysis revealed that polarity scores of on a Bard was significantly greater than ChatGPT (mean 0.16 vs. 0.11 on scale of -1 (negative) to 1 (positive); p=0.0323) and can be classified as "positive"; whereas subjectivity scores were similar across LLM's (mean 0.54 vs 0.50 on a scale of 0 (objective) to 1 (subjective), p=0.3030). Even though the majority of the LLM responses were appropriate, at this stage these chatbots should be considered as a versatile clinical resource to assist communication between clinicians and patients, and not a replacement of essential pre-anesthesia consultation. Further efforts are needed to incorporate health literacy that will improve patient-clinical communications and ultimately, post-operative patient outcomes.

Introduction

Pre-anesthesia evaluation is typically carried out 48-72 hours prior to the surgical procedure that requires administration of anesthesia. During this session, the anesthesiologists clarify numerous queries and misconceptions about anesthesia. Pre-operative instructions related to anesthesia can help alleviate post-operative recovery and can potentially help reduce patient's anxiety, length of stay, and post-operative pain [1]. In fact, a systematic review found that majority of the discussion with patients have been focused on preoperative anesthesia planning, and not much details on postoperative critical care communications exist [2]. Hence, patient-anesthesiologist communication is critical for simple tasks such as imparting factual information and explaining simple anesthesia related concepts. Now the question is, can we use artificial intelligence tools to answer these critical pre-anesthesia questions?

We aimed to evaluate if the responses generated from popular large language models (LLMs), a type of Generative Artificial Intelligence that generates text-based contents, such as OpenAI ChatGPT [3] and Google Bard [4] would be comparable to a anesthesiologist's response. Multidomain knowledge comprehension and problem-solving ability has been postulated as a specialty of theses LLM's. Even though ChatGPT has aced several medical board examination questions [5-11] and was deemed more "empathetic" as compared to clinicians [12], some of the studies revealed that LLMs are not widely accepted by clinicians for consultation purposes and has received a lot of negative attention since their debut [13]. These advanced computing technologies have gained popularity due to their ease of use and applicability to cross-functional fields across the globe. LLMs like ChatGPT or Bard work best by a process of prompt engineering – a set of specific instructions that are provided to LLMs to elicit specific and relevant responses [14]. Based on the level of instructions provided, the prompts can be classified as *zero-shot* (minimal instructions are provided), *few-shot* (instructions are for input and output are provided with examples), and the more complex *chain-of-thought* (complex reasoning and detailed instructions are provided) [14, 15]. In this study, we focused on the *zero-shot* approach and evaluated the LLM responses with minimal instructions. Prior studies have similarly utilized zero-shot queries from patient's perspective for other areas of medicine, [16-18] and primarily evaluated the AI generated responses through qualitative evaluations by clinicians. Our study aims to analyze the LLMs' response to patient's anesthesia-related queries in a more objective and quantitative manner. We have compared readability [16, 19-22], linguistic [23-26], and text-based analytical [13] differences between the two popular generative AI interfaces – ChatGPT and Bard.

Methods

Typical questions from patient's prior during pre-anesthesia sessions

We selected commonly asked anesthesia related questions during the pre-anesthesia consultation procedure prior to surgery in a hospital setting. Table 1 provides the list of questions that were analyzed in this study:

Table 1. Anesthesia related questions commonly asked by patients during pre-anesthesia evaluation

Response generation from large language models

We used the ChatGPT February 13, 2023 version (ChatGPT(F)), or ChatGPT March 22, 2023 version (ChatGPT(M)) (OpenAI, San Francisco)[3] and experimental version of Google Bard March 23, 2023 version (Google, Mountain View, CA) [4] in this study. *Zero-shot* prompt were provided (minimal instructions to the LLMs [6, 27-30]) to the questions listed in Table 1. To have a standardized response across platforms/versions, we added a prefixed statement to each question – "*Focus on patients' self-perceptions in preparation for anesthesia*". The generated responses were saved in a .txt file, and further evaluated for quantitative analyses. These LLMs are constantly updated and hence, the results are interpreted based on the versions used on the day of assessment. All queries were repeated three times to reduce variability [31](total of 33 queries each for ChatGPT(M) (uses Generative Pre-trained Transformer architecture (GPT-3.5)) and Bard (uses LaMDA (Language Model for Dialogue Applications) architecture); only 7 queries for ChatGPT(F) due to limited access and software upgrades).

Textual Response Evaluation

Detailed quantitative analysis for ChatGPT(F), ChatGPT(M), or Bard generated responses were evaluated using the metrics listed in Table 2.

Table 2. Metrics for analysis of responses generated by ChatGPT and Bard.

"Hallucinations"

"Hallucinations" are defined as errors in response from LLMs and these errors are framed by convincing statements that are not true. For this study, we are focusing on medical and surgery related facts. Factual accuracy of all LLM generated responses was qualitatively assessed by a clinician (UH). "Hallucination" error counts were expressed as a percentage of total responses (proportion).

Readability assessment

Readability is the basic understanding of the text that is expressed in the form of mathematical formulas or indices. For each generated response, Flesch reading ease (FRE) (Eq.1), Flesch-Kincaid grade level scoring (FKG) (Eq.2), and Simple Measure of Gobbledygook (SMOG) (Eq.3) readability assessments were performed using the online resource - TextCompare.org [16, 32-34]. The quantification of textual readability is performed as follows:

$$FRE = 206.835 - 1.1015 \times \left(\frac{words}{sentences}\right) - 84.6 \times \left(\frac{syllables}{words}\right) \quad \text{(Eq.1)[19]}$$

$$FKG = 0.39 \times \left(\frac{words}{sentences}\right) + 11.8 \times \left(\frac{syllables}{words}\right) - 15.59 \quad \text{(Eq.2)[21]}$$

$$SMOG = 3.1291 + 1.0430\sqrt{polysyllable\ count} \quad \text{(Eq.3)[20]}$$

where polysyllable count is the number of words of three or more syllables per 30 sentences.

FRE, FKG, and SMOG scores were converted to grade equivalents as per table provided in Table S1(a-c) (supplementary materials). These assessments calculate how difficult the

passage is to understand. For example, a high score on the Flesch ease test should have a corresponding low score on the Flesh-Kincaid grade-level test. Along the same lines, lower SMOG scores indicate high readability of the text. For patients, National Institutes of Health (NIH) recommends that the reading materials should be at a sixth grade reading level, whereas The American Medical Association (AMA) recommends an eight grade reading level [22, 35, 36]. Likewise, the NIH recommends the SMOG assessment for evaluating patient materials for clear and concise communication [20, 37]. Additionally, word count (WC) for each generated text was also calculated using the same online tool.

Lexical Diversity measurements

The complexity of the text is defined here in terms of mathematical indicators named as measure of textual lexical diversity (MTLD). MTLD is defined as the mean length of sequential word strings that maintains its Type-Token Ratio (TTR) value (TTR is set to 0.72), where TTR is defined as a ratio of unique words (types) to the total number of words (token) in the text [23, 38]. We compared of the generated responses using an online MTLD tool developed by Reuneker et al. [26, 39]. Larger MTLD values indicate better linguistic capabilities and more complex writing abilities. In a clinical setting, MLTD measures have been previously utilized as biomarkers to distinguish written text from aphasia patients as compared to neurological intact ones [24].

Computational Sentiment Analysis and Negative Word Detection

We performed computational sentiment analysis based on a vocabulary-based quantification or lexicon-based approach using *textblob 0.17.1* library (natural-language processing (NLP) tool) in *Jupyter Lab* environment (Python Interface) [40-42]. Words were extracted from the generated AI text and scored based on prior set rules or sentiment lexicon dictionary. Cumulative polarity and subjectivity scores were computed for each response. Polarity score between 0-1 was classified as “positive”, 0 was “neutral”, and less than zero until -1 was “negative”. Subjectivity score less than 0.5 were “objective”, and greater than 0.5 were “subjective”. For example, a positive score is given to words like “good”, “best”, “excellent”, etc., whereas negative score is given to words like “bad”, “awful”, “pathetic”, etc. Factual information constitutes lower subjectivity values and personal opinions constitute higher subjectivity values. In addition, we used the same Python package to scan the LLM generated texts for commonly misunderstood words during pre-anesthetic consultations - “aspiration”, “allergy”, “anaphylaxis, “local anesthetic”, “sedation”, “try”, and “worry” [43, 44].

Effect of Iterations

Since each question was repeated three times, we wanted to evaluate if there were any variations in generated texts between ChatGPT and Bard. For this reason, we quantified the pairwise similarity between the generated texts were calculated using Levenshtein distance score via countwordsfree.com [45-48]. Levenshtein distance was introduced in 1966 [45] and has been adapted to measures how similar or dissimilar sentences are from each other. Larger distances mean sentences are very dissimilar from each other, and vice versa. Since ChatGPT(F) responses were available for only seven questions (without repetition), the comparison of the responses using Levenshtein distance was confined to ChatGPT(M) and Bard responses across the three iterations.

Statistical Analysis

Unless stated otherwise, all data were expressed as mean ± std. deviation. Normality distribution of the data was checked using Shapiro Wilks test. All parameters, except for "Hallucinations" count, negative word count, and Levenshtein distance, were compared across the three LLMs (ChatGPT(F), ChatGPT(M), and Bard) using Kruskal-Wallis test and Dunn's test for multiple comparisons. "Hallucination" and negative word counts were taken as proportion of whole count, and a z-score test of proportions was utilized to evaluate the differences between groups. Levenshtein distances between ChatGPT(M) and Bard were compared across three iterations (i.e., three repeats per query) using Kruskal-Wallis test and Dunn's test for multiple comparisons. We wanted to evaluate if the length of the response influenced the quantified metrics from each LLM. We performed Spearman correlation analysis between all the quantified parameters vs. WC to evaluate the effect of number of words (WC) on each of the quantified parameters. We also evaluated the association between the polarity and objectivity scores generated from computational sentiment analysis using Spearman's correlation analysis. All analyses were performed using GraphPad Prism ver 9.5.1 (GraphPad Software, San Diego, CA, USA,) and data was considered significant at $p<0.05$.

Results

Responses from LLMs

Representative screenshots of the response generated by ChatGPT and Bard to anesthesia related queries are shown in Figure 1.

Fig. 1 Exemplary response from ChatGPT (A) and Bard (B) to one of the patient queries related to anesthesia.

Detailed responses from all LLMs are provided in the S2(a-c) (Supplementary Materials). Qualitatively, responses from ChatGPT(F) or ChatGPT(M) were more refined, abstract, and adequate, as compared to Bard. Upon close examination, all the queries answered by ChatGPT versions were correct, whereas some of the Google Bard answers were incorrect ("hallucinations") (Fig. 2; more details in S3 (supplementary materials)). Overall, accuracy of ChatGPT(F) (ChatGPT(F) vs. Bard - *z score* = 3.39; p=0.0007) and ChatGPT (M) (ChatGPT(M) vs. Bard - *z score* = 5.94; p<0.00001) were far superior (0/33; 0% error) to Google Bard (10/33; 30.3% error).

Fig. 2 "Hallucinations" - exemplary errors from Google Bard in response to the anesthesia related queries.

Readability assessment

FKG grade level scores were significantly higher for ChatGPT (M) as compared to ChatGPT (F) and Bard (14.74 ± 2.72 vs.14.36 ± 2.97 vs 9.4 ± 1.97, p=<0.0001) (Fig. 3(A)). With respect to the grade levels of ChatGPT(F), ChatGPT(M), and Bard correspond to "College level" (difficult to read), "College level" (difficult to read), and "8th and 9th grade" (conversational English), respectively (Table S1(a)). Similarly, FRE scores were found to 30 ± 21.21, 32.76 ± 14.14, and 55.35 ± 11.57 for ChatGPT(M), ChatGPT(F), and Bard, respectively (p=<0.0001) (Fig. 3(B)).

Conforming to Flesh-Kincaid Reading ease (Table (S1(b)), the FRE scores are at "College" levels for both ChatGPT versions, and "10th, 11th & 12th Grade" for Bard, respectively. SMOG scores of ChatGPT(F) and ChatGPT (M) were significantly different from Bard, but not statistically different amongst each other (16.94 ± 2.24 vs. 16.56 ± 2.0 vs.12.42 ± 1.49; p = <0.0001) (Fig. 3(C)). Per the SMOG index conversion (Table. S1(c)), the ChatGPT versions correspond to 7th grade level (fairly easy to read) and Bard corresponds to 6th grade level (easy to read). In terms of word count (WC), Bard responses were shorter than ChatGPT(F) (167.1 ± 50.9 vs. 244.4 ± 34.72, p =0.0007) or ChatGPT (M) (167.1 ± 50.9 vs. 203.9 ± 37.64, p = 0.0052).

Fig. 3 Detailed comparison of readability, linguistic, and text-based analytical of generated responses across ChatGPT(F), ChatGPT (M), and Bard. (A) FRG level, (B) FKE score, (C) SMOG, (D) MTLD, (E) Polarity, and (F) Objectivity parameters are compared across the three groups. Statistical significances are denoted by * (p<0.05), ** (p<0.005), * (p<0.001), and **** (p<0.0001).**

Lexical diversity measures

MTLD quantification was found to be significantly different across the groups (p = 0.0037), but pairwise statistical differences were achieved only between ChatGPT(M) and Bard (66.16 ± 13.42 vs. 55.27 ± 11.76 vs. 58.35 ± 11.07) (Fig. 3(D)).

Iteration effect

Levenshtein distances comparison across the three iterations of each question was similar for ChatGPT(M) and Bard (p=0.9335). Levenshtein distances between ChatGPT(M) and Bard were 1358.0 ± 253.6, 1341.0 ± 155.5, and 1356.0 ± 166.5 for first, second, and third iterations of the queries, respectively.

Sentiment Analysis and Negative Words Detection

Polarity scores for Bard were significantly greater than the two ChatGPT versions (0.16 ± 0.11 vs. 0.09 ± 0.07 vs. 0.11 ± 0.09; p=0.0323) (Fig. 3(E)) and classified as "positive" for all LLMs. The subjectivity scores were similar across all three LLM's – ChatGPT(F) 0.50 ± 0.09 vs. ChatGPT(M) 0.50 ± 0.09 vs. Bard 0.54 ± 0.11 (p=0.3030) (Fig. 3(F)) and classified as mildly "subjective", even though statistically insignificant. The proportion of negative words for ChatGPT(F) were lower than ChatGPT(M) (0%(0/7) vs. 39.39% (13/33) ; z = -2.02, p = 0.04338); however, the proportion of negative words between ChatGPT(M) and Bard (39.39% (13/33) vs. 21.21% (7/33); z = 1.6, p = 0.1074), and ChatGPT(F) and Bard (0%(0/7) vs. 21.21% (7/33) ; z = -1.34, p = 0.1802), were similar.

Effect of response length on quantified parameters

Association of word count with all other parameters are reported in Table 3. For ChatGPT(F), WC had no associations with any of the readability, lexical diversity, or sentiment analysis parameters. For ChatGPT(M), WC exhibited significantly moderate, negative association with polarity (ρ = -0.52; p = 0.0020) and subjectivity (ρ = -0.49; p = 0.0038) scores, respectively. Further, the polarity and subjectivity scores from ChatGPT(M) showed significant moderate,

positive association with each other (ρ = 0.62; p = 0.0001). For Bard, WC exhibited significant, moderate negative association with MTLD only (ρ = -0.41; p = 0.0167).

Table 3. Spearman correlation analysis (coefficients and p-values) for pairwise analysis of quantified parameters in this study.

Discussion

LLMs in patient related queries

In this study, we (i) compared the error ("hallucination") generated from LLMs when subjected to patient centric queries, (ii) showed that LLM's are prompt dependent, (iii) introduced textual data assessment metrics for objective evaluation of LLM responses, and (iv) estimated the feasibility of LLM becoming a futuristic healthcare tool. We aimed to explore the LLM's capability for future healthcare applications and more importantly if these AI-based chatbots are fit to answer patient's queries accurately. Our goal was to perform a data analytics based quantitative assessment of the responses (from ChatGPT and Bard) to queries from patients' point of view that were focused on anesthesia. Overall, ChatGPT has been more extensively tested for healthcare related and board exam questions as compared to Bard [5-8, 11, 16-18, 30, 34, 49-51]. To the best of our knowledge, this is the first study to quantitatively compare responses to anesthesia related questions between ChatGPT and Bard. Eleven common questions focusing on anesthesia were queried in chat interface of ChatGPT and Bard. The responses from these LLMs were then analyzed for word count, readability, linguistic quality (lexical diversity), sentiment analysis (polarity and subjectivity), and count of negative words commonly associated with poor surgical outcomes. A comparative account of the findings in this study and published literature are shown in Table 4.

Table 4. Comparative account of study parameters evaluated on ChatGPT and Bard responses [16, 31, 42, 49, 52-55].

ChatGPT outscored Bard in "Hallucinations"

Convincing, factually inaccurate responses with false claims (or "hallucinations") is a common occurrence in LLM responses [56-59]. In the research arena, the most common artifact is fabricated references that clearly do not exist. One example would be the investigation by Grigio et al. [60] where they queried ChatGPT for the use of anti-emetic drug, Olanzapine, for postoperative nausea and vomiting. ChatGPT generated four references in support of the Olanzapine query; however, none of the references generated were found in research databases. To the best of our knowledge, this is the first study to report "hallucinations" from Bard in response to anesthesia related queries (Fig.2). We found that ChatGPT, overall, has the upper hand when it comes to responses to patient queries and did not provide any incorrect answers or hallucinations, as opposed to Bard. The potential application of large language models such as ChatGPT for answering presurgical anesthesia related questions was suggested by Bhattacharya et al. [56]; however, their focus was more surgical oriented and theoretical in nature. Angel et al.[7] compared the responses from GPT-3, GPT-4, and Bard from

"Anesthesia Review: 1000 Questions and Answers to Blast the BASICS and Ace the ADVANCED" question bank and found that GPT-4 had the highest scores (78.33%) followed by GPT-3 (58.33%) and Bard (46.67%). Shay et al.[11] reported 56% success in ChatGPT responses to question from "Anesthesiology Examination and Board Review" book. Next, Aldridge et al. [5] reported moderate efficiency (63.6% accurate) of ChatGPT in answering Royal College of Anaesthetists fellowship exam questions (FRCA). Similarly, Birkett et al.[8] queried FRCA multiple choice questions and found that ChatGPT underperformed (proportion of correct responses was 0.697) and fell short of the official pass marks. Similar to our findings, a study comparing GPT-4, ChatGPT-3.5, and Bard performed on a neurosurgery oral boards preparatory question bank showed that Bard showed lowest performance score with accuracy of 44.2% (and highest "hallucination" rate), followed by ChatGPT-3.5 (62.4%) and GPT-4 (82.6%) [6]. Contrary to previous findings [5-7, 11], the LLM "hallucinations" was 0% for ChatGPT and 30.3% for Bard for anesthesia related queries. Concurring with previous studies [6, 7], we also report that responses from both ChatGPT versions were far superior and more accurate than Bard.

Readability of Bard was better than ChatGPT

Bard was less "wordier" compared to both ChatGPT versions (167.1 ± 50.9 vs. 244.4 ± 34.72 vs. 203.9 ± 37.64), and its responses in our study were comparable to ChatGPT studies by Johnson et al. (172.62 ± 34.77) [16] and Lee et al. (175.1 ± 58.2) [54]. Readability analysis of Bard responses using FKG, FRE, and SMOG were found to be at "8^{th} and 9^{th} grade", "10^{th}, 11^{th} & 12^{th} Grade", and "easy to read", respectively (Table 4). For Bard, the FKG levels align with the AMA level requirements [22, 35, 36] and the SMOG score concurs with NIH level requirements [20, 37], but the FRE is slight greater than both established norms. For ChatGPT(F) and ChatGPT(M), the readability for FKG and FRE was "college level" and the SMOG scores were "fairly easy to read". Like our study, Johnson et al. [16] addressed similar addressed cancer related myths and misconceptions using ChatGPT and found that there were noticeable higher FKG readability scores for ChatGPT (15.38 ± 3.28; "college level") as compared to National Cancer Institute's (NCI's) answers (12.04 ± 2.42; 10^{th}, 11^{th} & 12^{th} Grade"). The only other study comparing readability of ChatGPT and Bard responses was reported by Doshi et al. [49] and they investigated FKG score of radiological report simplification. Comparable to our study, they reported that Bard showed readability of "7^{th} grade level", whereas ChatGPT showed readability of "8^{th}-9^{th} grade level". A study by Ali et al. [52] was the only one that compared the readability of three tools, FKG, FKE, and SMOG, for evaluating ChatGPT responses for generating clinical letters. Contrary to our investigation, they reported "8^{th}-9^{th} grade level", "10th, 11th & 12th Grade", and "easy to read", for FKG, FKE, and SMOG scores, respectively. Several other studies focusing on ChatGPT are either better [55] or similar [31, 54] readability scores as compared to our findings, are listed in Table 4. Differences in readability measures between ChatGPT and Bard stem from their original computational architecture. Bard is developed on LaMDA (Language Model for Dialogue Applications), trained on 137 billion parameters, and it fetches data from the internet for each query. Whereas ChatGPT is powered by Generative Pre-trained Transformer (GPT), trained on 175 billion parameters, and is designed to possess natural language understanding, reasoning ability, and can generate coherent text in response to prompts. Since Bard was designed as a service-oriented tool, it thus possesses the ability to use simple words and craft sentences that is easily "readable".

LLM's are prompt dependent

Prompt engineering has opened up a wide arena for LLMs [15], and is considered as a backbone of many AI based tool today. Nastasi et al. [50] evaluated ChatGPT's responses from patients' perspective on acute illness, and queried critical questions concerning race, gender, and insurance status, and found the responses to be unsuited for personalized medical advice. They found ChatGPT changed clinical recommendations when social factors such as race or insurance status were modified. For this reason, we focused on a *zero-shot* learning approach that provides direct queries and minimal instructions to LLM to generate a response. Aldridge et al. [5] queried Royal College of Anaesthetists fellowship exam questions with GPT-3.5 and GPT-4, and found that ChatGPT(M) (utilizing GPT-4 technology) was able to answer questions with higher accuracy and exhibited a learning effect when subjected to repeated queries. To minimize variability, we repeated our queries three times, non-sequentially and found that pairwise Levenshtein distance scores to find similarities between the responses across the two LLMs (refer to *Iteration effect* in "Results" section). Thus, the generated responses from LLMs are user-centric, prompt-dependent, and their performance improves with repetition of queries.

ChatGPT exhibited more lexical diversity than Bard

We introduced a linguistic analysis for evaluating responses from LLMs, which has been previously utilized as a biomarker for aphasia patients [23]. MTLD is a robust lexical diversity metric and has been shown to be impartial to word count for a range of vocabularies and word counts/tokens [25]. A study by Herbold et al.[53] performed MTLD based lexical diversity analysis of essays written by human authors, ChatGPT-3, and ChatGPT-4, and found that ChatGPT-4 has more complex and diverse vocabulary compared to human authors or ChatGPT3 (108.91 ± 20.73 vs. 95.72 ± 23.50 vs 75.68 ± 12.89, respectively). Along the same lines, both ChatGPT versions were not influenced by word counts, and were greater in lexical diversity than Bard. Furthermore, Bard showed a negative association with increasing word count, and this could be since Bard's framework is designed to "conversational" and generates more human-like conversation, as compared to ChatGPT.

Sentiment analysis and negative word detection

Sentiment analysis quantification is typically utilized for classifying public sentiment in social media platforms, reviews, etc. We applied the same technique to LLM responses and quantified the generated responses for polarity (positive, neutral, or negative) and objectivity (objective or subjective) score quantification. Close to Chervenak et al. [42], we found that there was no difference across ChatGPT and Bard in the overall objectivity scores – meaning that LLM responses are typically indifferent to queries. However, we did find that Bard was more "positive" as compared to the two ChatGPT versions. This is possibly since Bard utilized NLP techniques to stimulate "human" like conversations and on the other hand, ChatGPT is designed to be better at "summarization" of complicated healthcare data [30, 42, 49, 52, 55].

Limitations and Future work

Variability in responses from ChatGPT and Bard across users must be addressed. Incomplete or inappropriate responses are quite common in LLM related queries. We repeated the questions three times and found no difference between the overall text outputs (Levenshtein distance metric). It is possible that repeating the queries can improve the overall performance of LLMs, as reported previously [5]. However, studies have utilized various number of iterations and there is no objective way to narrow down a specific iteration number that would work best for all scenarios. We have only utilized the free version of ChatGPT (GPT-3.5 based) in this

study and did not find many differences in the quantified metrics between the two versions of ChatGPT as access to the older version was limited due to software upgrades. Further, it is possible that there may be differences in responses to the queries in the paid version of ChatGPT (GPT-4 based). As noted by Nastasi et al. [50], there were noted differences in LLM response when the prompt included patient socioeconomic conditions, health insurance status, etc. Hence, more studies should focus on patient based LLM responses to include patient social determinants of health in future studies. Differences in initial prompts (like the *prefixed statement* utilized in this study) can impact the overall performance and accuracy of ChatGPT responses [6]. To take a step further, future studies can focus on level of complexity of the queries by incorporating the *Chain-of-Thought Prompting* (a step-by-step reasoning prior to answering questions)[29], as compared to *zero shot* (text-based instructions – this study) and/or *few-shot* approaches (text-based instructions with examples of input-output provided)[27, 28]. In addition to the response evaluation methods used in this study, several text-based quantifications such as automated evaluation metrics (ROUGE-L , METEOR, etc.), relationship extraction, dependency parsing, and intent extraction can be explored as well [30, 61-63]. The "hallucination" phenomenon of ChatGPT is a well-known problem and has been reported by several publications [56-59], and in fact, we show the Bard showed "hallucinations" in its responses as well. There is always a risk of fabricated evidence [60] to support a "hallucination" by LLMs, and that is not acceptable when it comes to communication of facts to a patients. The most important question – does the LLM "understand" the query given by the user like a human does? That question is still not clear. The development of LLMs for future healthcare tools would require additional layers of scrutiny and comprehension before it can be used for mainstream healthcare practices.

Conclusion

Our study shows that LLMs can generate effective responses from patient queries. ChatGPT was technical, precise, and descriptive, whereas Bard was conversational, adequate and exhibited "hallucinations". The best utilization of LLMs in a patient centric scenario would be for generating text for effective patient communication prior to surgery (this study), summarizing concise radiological report [49, 55], and many others faceted improvements in quality of perianesthesia patient care [64, 65]. Restriction of LLMs is a mere stopgap solution and more efforts are needed to integrate with current technology for improvement of patient care and advancement of medical research. It is important to note that LLM's cannot fully "think" and distinguish correction information from misinformation (yet). Creativity and ethical judgement are the hallmarks of a clinician, and that cannot be replaced with any technology.

References

1. Klafta, J.M. and M.F. Roizen, *Current understanding of patients' attitudes toward and preparation for anesthesia: a review.* Anesth Analg, 1996. **83**(6): p. 1314-21.
2. Tylee, M.J., et al., *Anesthesiologist to Patient Communication: A Systematic Review.* JAMA Netw Open, 2020. **3**(11): p. e2023503.
3. ChatGPT. *ChatGPT (Mar 23, 2023 version) [Large language model].* https://chat.openai.com/chat. 2023.
4. Bard. *Google Bard (Experimental version) [Large language model].* https://bard.google.com/. 2023.
5. Aldridge, M.J. and R. Penders, *Artificial intelligence and anaesthesia examinations: exploring ChatGPT as a prelude to the future.* Br J Anaesth, 2023.
6. Ali, R., et al., *Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank.* Neurosurgery, 2023: p. 2023.04.06.23288265.
7. Angel, M.C., et al., *Clinical Knowledge and Reasoning Abilities of AI Large Language Models in Anesthesiology: A Comparative Study on the ABA Exam.* medRxiv, 2023: p. 2023.05.10.23289805.
8. Birkett, L., T. Fowler, and S. Pullen, *Performance of ChatGPT on a primary FRCA multiple choice question bank.* Br J Anaesth, 2023.
9. Johnson, D., et al., *Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model.* Res Sq, 2023.
10. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.* PLOS Digit Health, 2023. **2**(2): p. e0000198.
11. Shay, D., et al., *Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions.* Br J Anaesth, 2023.
12. Ayers, J.W., et al., *Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum.* JAMA Intern Med, 2023. **183**(6): p. 589-596.
13. Praveen, S.V. and V. Vajrobol, *Can ChatGPT be Trusted for Consulting? Uncovering Doctor's Perceptions Using Deep Learning Techniques.* Ann Biomed Eng, 2023.
14. Radford, A. and K. Narasimhan. *Improving Language Understanding by Generative Pre-Training*. 2018.
15. Wei, J., et al., *Chain of thought prompting elicits reasoning in large language models.* arXiv preprint arXiv:2201.11903, 2022.
16. Johnson, S.B., et al., *Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information.* JNCI Cancer Spectr, 2023. **7**(2).
17. Sarraju, A., et al., *Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model.* JAMA, 2023. **329**(10): p. 842-844.
18. Yeo, Y.H., et al., *Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma.* medRxiv, 2023: p. 2023.02.06.23285449.
19. Flesch, R., *A new readability yardstick.* J Appl Psychol, 1948. **32**(3): p. 221-33.
20. McLaughlin, G.H., *SMOG Grading-a New Readability Formula.* Journal of Reading, 1969. **12**(8): p. 639-646.
21. Kincaid, J.P., et al., *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. 1975, Naval Technical Training Command Millington TN Research Branch.
22. Rooney, M.K., et al., *Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis.* J Patient Exp, 2021. **8**: p. 2374373521998847.

23. McCarthy, P.M. and S. Jarvis, *MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment.* Behav Res Methods, 2010. **42**(2): p. 381-92.
24. Fergadiotis, G. and H.H. Wright, *Lexical diversity for adults with and without aphasia across discourse elicitation tasks.* Aphasiology, 2011. **25**(11): p. 1414-1430.
25. Koizumi, R. and Y. In'nami, *Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens.* System, 2012. **40**(4): p. 554-564.
26. Reuneker, A. *Measures of Lexical Diversity (LD). Retrieved April 20, 2023, from https://www.reuneker.nl/files/ld/.* 2023.
27. Radford, A., et al. *Language Models are Unsupervised Multitask Learners*. 2019. OpenAI Blog,1(8):9.
28. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
29. Qin, C., et al., *Is chatgpt a general-purpose natural language processing task solver?* arXiv preprint arXiv:2302.06476, 2023.
30. Tang, L., et al., *Evaluating Large Language Models on Medical Evidence Summarization.* medRxiv, 2023: p. 2023.04.22.23288967.
31. Momenaei, B., et al., *Appropriateness and Readability of ChatGPT-4 generated Responses for Surgical Treatment of Retinal Diseases.* Ophthalmol Retina, 2023.
32. TextCompare. *https://www.textcompare.org/readability/ Accessed on March 22, 2023*. 2023.
33. Sinha, C., et al., *Qualitative Exploration Toward the Development of a Parent-Rated Scale for Insomnia in Children with Autism Spectrum Disorder.* J Autism Dev Disord, 2023.
34. Juhi, A., et al., *The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions.* Cureus, 2023. **15**(3): p. e36272.
35. Weis, B., *Health Literacy: A Manual for Clinicians. Chicago, IL: American Medical Association, American Medical Foundation; 2003.• National Institutes of Health. How to Write Easy to Read Health Materials: National Library of Medicine Website.* How to Write Easy to Read Health Materials: National Library of Medicine Website, 2003.
36. Health, N.I.o. *MedlinePlus: How to write easy to read health materials. https://medlineplus.gov/ency/patientinstructions/000455.htm*. 2023.
37. Health, N.I.o. *Clear Communication Guide. https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple*. 2023.
38. Association, A.E.R., *Standards for educational and psychological testing*. 1999: American Educational Research Association.
39. Matthews, T., *Recreating Literary Works Using ChatGPT and Evaluating Results with NLP Analysis. https://rpubs.com/datadataguy13/1011253*. 2023.
40. Steven Loria, P.K., Matthew Honnibal, Roman Yankovsky, David Karesh, Evan Dempsey, Wesley Childs, Jeff Schnurr, Adel Qalieh, Lage Ragnarsson, Jonathon Coe, Adrián López Calvo, Nitish Kulshrestha, Jhon Eslava, Tyler James Harden, Jeff Kolb, Daniel Ong, Jamie Moschella, Roman Korolev, Ram Rachum, Romain Casati, Evgeny Kemerov, *textblob 0.17.1 https://textblob.readthedocs.io/.* 2013.
41. Thomas Kluyver, B.R.-K., Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, Jupyter Development Team, *Jupyter Notebooks – a publishing format for reproducible computational workflows*, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas (pp. 87–90)*, I.F.L.B.S. (Eds.), Editor. 2016.

42. Chervenak, J., et al., *The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations.* Fertil Steril, 2023.
43. Babitu, U.Q. and A.M. Cyna, *Patients' understanding of technical terms used during the pre-anaesthetic consultation.* Anaesth Intensive Care, 2010. **38**(2): p. 349-53.
44. Cyna, A.M., *A GREAT interaction and the LAURS of communication in anesthesia.* Acta Anaesthesiologica Belgica, 2018. **69**(3): p. 131-135.
45. Levenshtein, V.I. *Binary codes capable of correcting deletions, insertions, and reversals*. in *Soviet physics doklady*. 1966.
46. Efremov, E., Vassilev, V. . *Acurracy of ASR Software in Bulgarian Language and its Application in Journalism*. in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 2022.
47. Sanchez-Pizani, R., et al., *Hybrid flexible (HyFlex) seminar delivery – A technical overview of the implementation.* Building and Environment, 2022. **216**: p. 109001.
48. Countwordsfree. *https://countwordsfree.com/comparetexts/ Accessed on March 22, 2023*. 2023.
49. Doshi, R., et al., *Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing.* medRxiv, 2023: p. 2023.06.04.23290786.
50. Nastasi, A., et al., *Does ChatGPT Provide Appropriate and Equitable Medical Advice?: A Vignette-Based, Clinical Evaluation Across Care Contexts*. 2023, medRxiv.
51. Van Bulck, L. and P. Moons, *What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value and danger of ChatGPT-generated responses to health questions.* Eur J Cardiovasc Nurs, 2023: p. zvad038.
52. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters.* Lancet Digit Health, 2023. **5**(4): p. e179-e181.
53. Herbold, S., et al., *AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays.* arXiv preprint arXiv:2304.14276, 2023.
54. Lee, T.C., et al., *ChatGPT Answers Common Patient Questions About Colonoscopy.* Gastroenterology, 2023.
55. Li, H., et al., *Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports.* Clinical Imaging, 2023.
56. Bhattacharya, K., et al., *ChatGPT in Surgical Practice-a New Kid on the Block.* Indian Journal of Surgery, 2023.
57. Alkaissi, H. and S.I. McFarlane, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing.* Cureus, 2023. **15**(2): p. e35179.
58. Azamfirei, R., S.R. Kudchadkar, and J. Fackler, *Large language models and the perils of their hallucinations.* Crit Care, 2023. **27**(1): p. 120.
59. Liu, S., et al., *Assessing the Value of ChatGPT for Clinical Decision Support Optimization.* medRxiv, 2023: p. 2023.02.21.23286254.
60. Grigio, T.R., H. Timmerman, and A.P. Wolff, *ChatGPT in anaesthesia research: risk of fabrication in literature searches.* Br J Anaesth, 2023.
61. Mamou, J., et al., *Term set expansion based nlp architect by intel ai lab.* arXiv preprint arXiv:1808.08953, 2018.
62. Chen, X., et al., *How Robust is GPT-3.5 to Predecessors? A Comprehensive Study on Language Understanding Tasks.* arXiv preprint arXiv:2303.00293, 2023.
63. Wang, J., et al., *On the robustness of chatgpt: An adversarial and out-of-distribution perspective.* arXiv preprint arXiv:2302.12095, 2023.
64. Odom-Forren, J., *The Role of ChatGPT in Perianesthesia Nursing.* J Perianesth Nurs, 2023. **38**(2): p. 176-177.

65. Tewfik, G., et al., *ChatGPT and Its Potential Implications for Clinical Practice: An Anesthesiology Perspective.* Biomed Instrum Technol, 2023. **57**(1): p. 26-30.

Comparison of ChatGPT vs. Bard to Anesthesia-related Queries

Sourav S. Patnaik, Ulrike Hoffmann

Contents

Table 1. Anesthesia related questions commonly asked by patients during pre-anesthesia evaluation

Questions	[¥]ChatGPT February 13, 2023 version	ChatGPT March 14, 2023 version	Bard March 22, 2023 version
1. What are the most commonly asked question by patients regarding anesthesia?	✓	✓	✓
2. Will I ever wake up again?	✓	✓	✓
3. If I have multiple anesthetics, will I get brain damage?	✓	✓	✓
4. Will I lose my memory?	✓	✓	✓
5. Will I have nausea and vomiting when I wake up?	✓	✓	✓
6. How does anesthesia work?	✓	✓	✓
7. What words are most associated with fear of anesthesia?	✓	✓	✓
8. What does an anesthesiologist do?	×	✓	✓
9. What is intubation?	×	✓	✓
10. Will I be on a machine that breathes for me? (or something similar)	×	✓	✓
11. Will I be in pain after surgery?	×	✓	✓

[¥] Only seven queries could be completed prior to software upgrade

Table 2. Metrics for analysis of responses generated by ChatGPT and Bard.

Type of analysis	Metric	Interpretation	Score Ranges	Reference
Reading assessment	Flesch Kincaid Grade Level (FKG) Flesch-Kincaid readability reading ease (FKE)	Uses word length and sentence lengths to calculate how difficult the passage is to understand. E.g., a high score on the ease test should have a corresponding low score on the Grade-level test.	FKG score ranges from 0-20+; FRE ranges from 0 to 100	[1-3]
	SMOG score	Primarily used in healthcare. Focuses on polysyllabic words - words with multiple syllables	0 to 211 or higher	
Lexical diversity measures	Measure of textual lexical diversity (MTLD)	Measure of the range of different words in the text	Varies; lowest = 0	[4, 5]
Computational Sentiment analysis	Polarity score	It indicates the sentiment polarity or sentiment orientation of a given text. It measures the degree of positivity, negativity, or neutrality in the text.	A normalized score of -1 to +1 is calculated for the text; 0 is designated as “neutral”, between 0 to -1 is “negative”, and between 0 to 1 is “positive”	[6, 7]
	Subjectivity score	Subjectivity refers to the presence of personal opinions, beliefs, or emotions, while objectivity refers to the absence of personal bias or emotions, focusing more on factual information.	Subjectivity ranges from 0 to 1, with 0 being “objective” and anything >0.5 being “subjective”	
Text similarity	Levenshtein distance	It’s a measure of how different the two strings are or their linguistic distance.	Varies; lowest = 0	[8, 9]
Others	Word counts (WC)	Total numbers of words in the text	lowest = 0	
	Negative words count	Count of negative words associated with poor	lowest = 0	[10, 11]

		outcomes in pre-anesthesia consultation (-“aspiration”, “allergy”, “anaphylaxis, “local anesthetic”, “sedation”, “try”, and “worry”)		

Table 3. Association between variables quantified in this study. Bold numbers are significantly different p<0.05.

Correlation Analysis		ChatGPT(F) (n=7)		ChatGPT(M) (n=33)		Bard (n=33)	
		ρ	*p-value*	ρ	*p-value*	ρ	*p-value*
WC	FKG	0.59	0.1698	0.21	0.2415	0.15	0.4193
WC	FKE	-0.57	0.2000	-0.26	0.1399	-0.28	0.1212
WC	SMOG	0.68	0.1095	0.20	0.2583	0.16	0.3713
WC	MTLD	0.29	0.5560	-0.06	0.7274	***-0.41***	***0.0167***
WC	Polarity	-0.29	0.5286	***-0.52***	***0.0020***	-0.30	0.0860
WC	Subjectivity	-0.56	0.2056	***-0.49***	***0.0038***	-0.33	0.0591
Polarity	Subjectivity	0.18	0.6972	***0.62***	***0.0001***	0.26	0.1434

Table 4. Comparative account of study parameters evaluated on ChatGPT and Bard responses

Parameter	ChatGPT	Bard	Information	Source
FKG	14.36 ± 2.97[¥] 14.74 ± 2.72[£]	9.4 ± 1.97	Anesthesia related queries	This study
	9.81	-	Generated clinical letters	[12]
	9[€] 9.3[£]	7.9	Radiology report simplification (prompt 1)[¶]	[13]
	5.8 ± 1.1	-	Diagnostic radiology report simplification	[14]
	15.38 ± 3.28	-	Cancer myths and misconceptions	[15]
	14.1 ± 2.6 14 ± 1.3 14.8 ± 1.3	-	[£]Response to questions related to surgical treatment of retinal diseases - retinal detachment - macular hole - epiretinal membrane	[16]
	13.1 ± 2.2	-	Patient questions about colonoscopy	[17]
FRE	32.76 ± 14.14[¥] 30 ± 21.21[£]	55.35 ± 11.57	Anesthesia related queries	This study
	55.63	-	Generated clinical letters	[12]
	83.5 ± 5.6	-	Diagnostic radiology report simplification	[14]
	32.3 ± 10.8 34.4 ± 7.7 28.1 ± 7.5	-	[£]Response to questions related to surgical treatment of retinal diseases - retinal detachment - macular hole epiretinal membrane	[16]
SMOG	16.94 ± 2.24[¥] 16.56 ± 2.0[£]	12.42 ± 1.49	Anesthesia related queries	This study
	12.50	-	Generated clinical letters	[12]

	16.4 ± 2.4 16.2 ± 1.3 16.4 ± 1.4	-	£Response to questions related to surgical treatment of retinal diseases - retinal detachment - macular hole epiretinal membrane	[16]
MTLD	58.35 ± 11.07¥ 66.16 ± 13.42£	55.27 ± 11.76	Anesthesia related queries	This study
	75.68 ± 12.89¥ 108.91 ± 20.73£	-	Essay writing	[18]
Polarity	0.09 ± 0.07¥ 0.11 ± 0.09£	0.16 ± 0.11	Anesthesia related queries	This study
	0.11¥	-	Fertility related queries	[19]
Subjectivity	0.50 ± 0.09¥ 0.50 ± 0.09£	0.54 ± 0.11	Anesthesia related queries	This study
	0.42¥	-	Fertility related queries	[19]

¶Median values
¥ChatGPT February 13 version (ChatGPT(F))
€ChatGPT-3.5 or ChatGPT March 23 version (ChatGPT(M))
£ChatGPT-4 (paid version)

References

1. Health, N.I.o. *Clear Communication Guide.* https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple. 2023.
2. McLaughlin, G.H., *SMOG Grading-a New Readability Formula.* Journal of Reading, 1969. **12**(8): p. 639-646.
3. Fitzsimmons, P.R., et al., *A readability assessment of online Parkinson's disease information.* J R Coll Physicians Edinb, 2010. **40**(4): p. 292-6.
4. McCarthy, P.M. and S. Jarvis, *MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment.* Behav Res Methods, 2010. **42**(2): p. 381-92.
5. Matthews, T., *Recreating Literary Works Using ChatGPT and Evaluating Results with NLP Analysis.* https://rpubs.com/datadataguy13/1011253. 2023.
6. Wiebe, J., T. Wilson, and C. Cardie, *Annotating Expressions of Opinions and Emotions in Language.* Language Resources and Evaluation, 2006. **39**(2-3): p. 165-210.
7. Lee, L. and B. Pang, *Opinion Mining and Sentiment Analysis.* Foundations and Trends® in Information Retrieval, 2008. **2**(1–2): p. 1-135.
8. Efremov, E., Vassilev, V. . *Acurracy of ASR Software in Bulgarian Language and its Application in Journalism*. in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 2022.
9. Sanchez-Pizani, R., et al., *Hybrid flexible (HyFlex) seminar delivery – A technical overview of the implementation.* Building and Environment, 2022. **216**: p. 109001.
10. Babitu, U.Q. and A.M. Cyna, *Patients' understanding of technical terms used during the pre-anaesthetic consultation.* Anaesth Intensive Care, 2010. **38**(2): p. 349-53.
11. Cyna, A.M., *A GREAT interaction and the LAURS of communication in anesthesia.* Acta Anaesthesiologica Belgica, 2018. **69**(3): p. 131-135.
12. Ali, S.R., et al., *Using ChatGPT to write patient clinic letters.* Lancet Digit Health, 2023. **5**(4): p. e179-e181.
13. Doshi, R., et al., *Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing.* medRxiv, 2023: p. 2023.06.04.23290786.
14. Li, H., et al., *Decoding radiology reports: Potential application of OpenAI ChatGPT to enhance patient understanding of diagnostic reports.* Clinical Imaging, 2023.
15. Johnson, S.B., et al., *Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information.* JNCI Cancer Spectr, 2023. **7**(2).
16. Momenaei, B., et al., *Appropriateness and Readability of ChatGPT-4 generated Responses for Surgical Treatment of Retinal Diseases.* Ophthalmol Retina, 2023.
17. Lee, T.C., et al., *ChatGPT Answers Common Patient Questions About Colonoscopy.* Gastroenterology, 2023.
18. Herbold, S., et al., *AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays.* arXiv preprint arXiv:2304.14276, 2023.
19. Chervenak, J., et al., *The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations.* Fertil Steril, 2023.

(A) ChatGPT Response

SO Q. Focus on patients' self-perceptions in preparation for anesthesia. What does an anesthesiologist do?

An anesthesiologist is a medical doctor who specializes in administering anesthesia to patients before, during, and after surgery or other medical procedures. The role of an anesthesiologist is to ensure that patients are safe, comfortable, and pain-free during these procedures.

Before a procedure, an anesthesiologist will review the patient's medical history and perform a physical exam to determine the most appropriate type and dose of anesthesia. They will also explain the anesthesia process to the patient, answer any questions they may have, and address any concerns or fears they may have about the procedure.

During the procedure, the anesthesiologist is responsible for monitoring the patient's vital signs, including heart rate, blood pressure, and oxygen levels, and adjusting the anesthesia as necessary to ensure the patient's safety and comfort. They may also manage other aspects of the patient's care, such as pain management and fluid balance.

After the procedure, the anesthesiologist will continue to monitor the patient's recovery and manage any postoperative pain or Regenerate response ensure that the patient is awake, alert, and able to breathe on their own before they are discharged from the hospital

Send a message...

ChatGPT Mar 23 Version. Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts.

(B) Bard Response

Focus on patients' self-perceptions in preparation for anesthesia. What does an anesthesiologist do?

View other drafts

An anesthesiologist is a doctor who specializes in anesthesia, which is the medical treatment that keeps you from feeling pain during procedures or surgery. Anesthesiologists have a wide range of responsibilities, including:

- Assessing your health and medical history before surgery to make sure that anesthesia is safe for you.
- Discussing your options for anesthesia with you and your surgeon.
- Administering anesthesia during surgery.
- Monitoring your vital signs and breathing during surgery.
- Helping you wake up after surgery and manage any pain you may have.

In addition to these medical responsibilities, anesthesiologists also play an important role in helping patients cope with the emotional and psychological aspects of surgery. Many patients experience anxiety or fear about surgery, and anesthesiologists can help to allay these fears by providing information and support. They can also help patients to develop coping mechanisms for dealing with pain and stress.

Enter a prompt here

Bard may display inaccurate or offensive information that doesn't represent Google's views.

Exemplary error #1

5. **What should I do to prepare for anesthesia?**

Here are some things you can do to prepare for anesthesia:

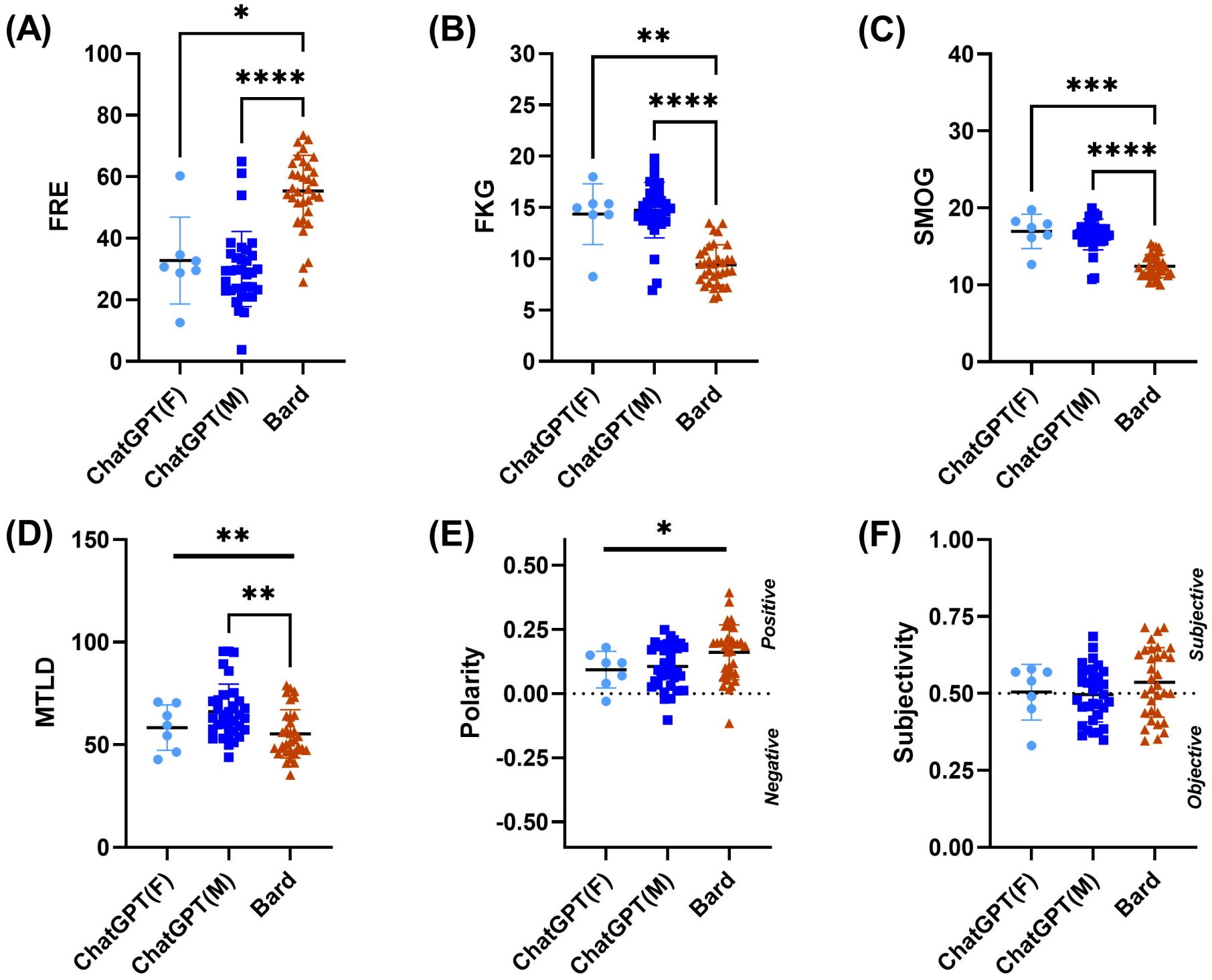
- Tell your doctor about all of your medications, including prescription and over-the-counter medications.
- Tell your doctor about any allergies you have.
- Stop smoking and drinking alcohol at least 24 hours before your surgery.
- Eat a light breakfast on the day of your surgery.

It is important to follow your doctor's instructions carefully before and after your surgery.

Exemplary error #2

Q10. Focus on patients' self-perceptions in preparation for anesthesia. Will I be on a machine that breathes for me? (or something similar)

Yes, you will be on a machine that breathes for you if you are unconscious during surgery. This is called mechanical ventilation. The machine will deliver oxygen to your lungs and remove carbon dioxide from your blood. You will not be able to breathe on your own while you are on the machine.


(A)
FRE
100
80
60
40
20
0
*
****
ChatGPT(F)
ChatGPT(M)
Bard
(B)
FKG
30
25
20
15
10
5
0
**
****
ChatGPT(F)
ChatGPT(M)
Bard
(C)
SMOG
40
30
20
10
0
***
****
ChatGPT(F)
ChatGPT(M)
Bard
(D)
MTLD
150
100
50
0
**
**
ChatGPT(F)
ChatGPT(M)
Bard
(E)
Polarity
0.50
0.25
0.00
-0.25
-0.50
*
Positive
Negative
ChatGPT(F)
ChatGPT(M)
Bard
(F)
Subjectivity
1.00
0.75
0.50
0.25
0.00
Subjective
Objective
ChatGPT(F)
ChatGPT(M)
Bard