J Med Edu. 2023 December; 22(1):e137753. https://doi.org/10.5812/jme-137753.

Published online 2023 October 17. **Research Article**

Employing Large Language Models for Surgical Education: An In-depth Analysis of ChatGPT-4

Adrian Hang Yue Siu [1, 2, *], Damien Gibson [3], Xin Mu [4], Ishith Seth [4, 5], Alexander Chi Wang Siu [6], Dilshad Dooreemeah [5] and Angus Lee [5]

[1] Surgical Outcomes Research Centre, Royal Prince Alfred Hospital, Sydney, Australia
[2] Faculty of Medicine and Health, Central Clinical School, The University of Sydney, Sydney, New South Wales, Australia
[3] Department of Surgery, St George Hospital, Sydney, Australia
[4] Department of Surgery, Peninsula Health, Victoria, Australia
[5] Department of Surgery, Bendigo Health, Victoria, Australia
[6] School of Medical Sciences, University of Sydney, Sydney, Australia

[*] *Corresponding author*: Research Affiliate, Surgical Outcomes Research Centre, Royal Prince Alfred Hospital, Sydney, Australia. Email: adriansiu7@hotmail.com

Received 2023 May 22; **Revised** 2023 September 13; **Accepted** 2023 September 17.

Abstract

Background: The growing interest in artificial intelligence (AI) has spurred an increase in the availability of Large Language Models (LLMs) in surgical education. These LLMs hold the potential to augment medical curricula for future healthcare professionals, facilitating engagement in remote learning experiences, and assisting in personalised student feedback.
Objectives: To evaluate the ability of LLMs to assist junior doctors in providing advice for common ward-based surgical scenarios with increasing complexity.
Methods: Utilising an instrumental case study approach, this study explored the potential of LLMs by comparing the responses of the ChatGPT-4, BingAI and BARD. LLMs were prompted by 3 common ward-based surgical scenarios and tasked with assisting junior doctors in clinical decision-making. The outputs were assessed by a panel of two senior surgeons with extensive experience in AI and education, qualitatively utilising a Likert scale on their accuracy, safety, and effectiveness to determine their viability as a synergistic tool in surgical education. A quantitative assessment of their reliability and readability was conducted using the DISCERN score and a set of reading scores, including the Flesch Reading Ease Score, Flesch-Kincaid Grade Level, and Coleman-Liau index.
Results: BARD proved superior in readability, with Flesch Reading Ease Score 50.13 (± 5.00), Flesch-Kincaid Grade Level 9.33 (± 0.76), and Coleman-Liau index 11.67 (± 0.58). ChatGPT-4 outperformed BARD and BingAI, with the highest DISCERN score of 71.7 (± 2.52). Using a Likert scale-based framework, the surgical expert panel further affirmed that the advice provided by the ChatGPT-4 was suitable and safe for first-year interns and residents. A *t*-test showed statistical significance in reliability among all three AIs ($P < 0.05$) and readability only between the ChatGPT-4 and BARD. This study underscores the potential for LLM integration in surgical education, particularly ChatGPT, in the provision of reliable and accurate information.
Conclusions: This study highlighted the potential of LLM, specifically ChatGPT-4, as a valuable educational resource for junior doctors. The findings are limited by the potential of non-generalizability of the use of junior doctors' simulated scenarios. Future work should aim to optimise learning experiences and better support surgical trainees. Particular attention should be paid to addressing the longitudinal impact of LLMs, refining AI models, validating AI content, and exploring technological amalgamations for improved outcomes.

Keywords: Surgical Education, Medical Education, Large Language Models, Artificial Intelligence, COVID-19 Pandemic, Decision Aids

1. Background

Chat Generative Pre-Trained Transformer 4 (ChatGPT-4) (Open AI), BARD (Google), and BingAI (Microsoft) are state-of-the-art large language models (LLM), that generate human-like language to answer questions and complete text (1). Currently, there is a growing prevalence of chatbots and artificial intelligence (AI) in daily life, from assisting with school homework to fooling researchers with phony abstracts (2) While discussions surrounding ownership, authorship, and potential misuse continue to be debated (3, 4), there has been a growing trend of artificial intelligence in medical education as a disruptive

technology (5-7).

Trained by a vast corpus of medical literature, these LLMs can provide students with detailed and relevant information on any chosen subject matter (8). ChatGPT-3.5 has demonstrated performance near the passing threshold of 60% accuracy in the USMLE Steps 1, 2 CK, and 3 exams, comparable to a first-year postgraduate doctor seeking licensure as an unsupervised physician in the United States of America (USA) (5). Following the March 2023 release of ChatGPT-4, it has been claimed that the updated version of this LLM has enhanced clinical reasoning and test-answering capabilities compared to previous iterations (9, 10). This has been further demonstrated with ChatGPT-4 significantly outperforming ChatGPT-3.5 on American neurosurgical written board examinations (8).

Concurrently, the COVID-19 pandemic has fundamentally reshaped medical education owing to public health concerns and stay-at-home mandates in Australia, leading to reduced face-to-face teaching and clinical exposure for medical students and junior doctors. This has triggered concerns regarding the potential long-term effects of medical training. However, the pandemic has also spurred innovations in medical education, particularly in virtual simulation and telehealth (5, 10, 11). While most universities resume face-to-face training, the growing prevalence of AI and LLMs cannot be ignored, as they have emerged as possible tools to aid informed diagnosis and make safer treatment plans. The unique ability of LLMs to process vast amounts of clinical data and current information positions them as theoretical adjuncts for medical students and junior doctors. However, the challenge therefore, is ensuring that their everyday use in medical education is not at the cost of critical thinking and clinical acumen. Additionally, given the prevalence of LLMs in the post-pandemic context, it must also be considered whether they have a role in virtual simulation and remote learning.

This study aimed to assess the potential of LLMs, with a focus on ChatGPT-4, BingAI, and BARD, to aid surgical education and offer reliable advice to junior doctors (post-graduate years one or two). Comparison of these LLMs will be conducted through comprehensive quantitative and qualitative assessment, which will provide valuable insights into the potential use of LLMs in surgical education. In particular, the limitations of AI and LLMs are also briefly explored to understand the boundaries of this technology. Ultimately, by exploring these issues in-depth, this study will help reshape the traditional medical education curriculum.

Using a case study approach, each LLM will be prompted in three routine ward-based surgical situations to aid clinical decision-making. These scenarios were formulated based on real-life clinical scenarios and textbook case studies (11, 12), with a final review by expert general surgeons. Responses were qualitatively evaluated for accuracy, safety, and effectiveness using a Likert scale (13). Their validity, reliability, and readability were quantitatively assessed using the DISCERN score (14) and various readability metrics, including the Flesch Reading Ease (FRE) Score (15), Flesch-Kincaid Grade Level (FKGL) (16), and Coleman-Liau index (CLI) (17). It is hypothesised that their outputs may exhibit disparities in reliability and readability owing to differing training data.

2. Methods

2.1. Study Design

To address the primary research aim, this study adopted an instrumental case study approach (18). This research method is often used to understand and gain insight into a phenomenon in a context, which in our case, is the use of AI LLMs (ChatGPT-4, BARD, and BingAI) in medical education.

2.2. Methodology

A series of three increasingly complex clinical scenarios were posed to AI LLMs (ChatGPT-4, BARD, and BingAI). These scenarios were common ward-based surgical reviews performed by a junior doctor. These scenarios were formulated and designed from real-life clinical scenarios that the authors had encountered as ward-based junior doctors. The scenarios were then validated for accuracy and relevance with textbook analysis of case studies (11, 12). These case studies were evaluated by a panel of two board-certified surgeons independently (AL and DD) with over 20 years of experience. Responses were qualitatively evaluated for accuracy, safety, and effectiveness using a Likert scale (13). If any differences in the Likert scale or reliability tools arose, these were discussed until consensus was achieved.

The responses from each scenario then underwent qualitative analysis for accuracy, appropriateness, and patient safety. The quantitative assessment standards comprised of two aspects: Reliability, which was determined using the DISCERN score (14), and readability, which was evaluated through three widely recognized scoring systems: FRE score (15), FKGL (16), and CLI (17). Due to differing training data, it is hypothesised that their outputs may exhibit disparities in reliability and readability.

2.3. Data Collection Tools

The Likert scale (13) used in this study, is a 5-point global scale to qualitatively evaluate the accuracy, safety and effectiveness of the three LLMs. The 5-point scale consisted of two utmost poles ('Strong Agree' and 'Strongly disagree') and neutral option ('Neither agree nor disagree'), linked with intermediate answer options ('Agree' and 'Disagree').

The DISCERN questionnaire (14) was used to quantitatively assess the reliability of written information from the LLM outputs, and is considered a valid and reliable score for evaluating consumer health information (19). According to the literature (20), DISCERN scores may be categorised as follows: 'excellent' for scores of 63 to 75 points, 'good' for scores of 51 to 62 points, 'fair' for scores of 39 to 50 points, 'poor' for scores of 27 to 38 points, and 'very poor' for scores of 16 to 26 points.

In this study, readability was assessed using three recognised scoring systems: Flesch Reading Ease Score (15), Flesch-Kincaid Grade Level (16), and Coleman-Liau index (17). The FRE score and FKGL are both calculated using the average sentence length (i.e., number of words divided by the number of sentences) and the average syllables per word (i.e., number of syllables divided by the number of words) (21). The CLI (17) is calculated using the average number of letters per 100 words, and the average sentence length. Scoring of the FRE is through a 100-point scale with higher scores indicating higher readability and easier to understand text. Alternatively, FKGL and CLI, indicate the USA academic grade level (number of years of education) necessary to comprehend the written material. These readability tests were selected due to their wide and validated use in previous studies (21, 22).

2.4. Expert Review Framework

To evaluate the AI outputs, a Likert scale-based framework was employed (Table 1). Each LLM output was assessed by a panel of expert General Surgeons using this framework. The panel of board-certified surgeons conceptualised the research idea and both expert surgeons were recruited due to their experience in AI, research, and education. The surgeons' credentials extended beyond medical degrees to include specialised surgical training, affiliations with professional medical bodies, and leading roles at esteemed medical institutions. Their proficiency in evaluating AI outputs, understanding of its implications in medical education, and previous experience with AI research projects substantiated their selection for the panel. Experts were asked to rate the accuracy, reliability, proficiency, comprehensiveness, relevancy, general knowledge, errors, citations, and references of AI-generated responses.

2.5. Statistical Analysis

Between the three LLM, a student's *t*-test (23) was conducted to determine the statistical significance of the reliability and readability scores. Further commentary and critique of the answers were provided by two specialist general surgeons with extensive clinical experience, who provided an expert review framework on the subject matter. Statistical analyses were then performed on the collected data to determine the AI's performance across different dimensions, with a focus on identifying areas of strength and potential improvement.

2.6. Selection of LLMs

Due to the probabilistic algorithm and random-sampling method of LLMs, answers may vary slightly even if the same question is asked. For this study, three scenarios were placed into ChatGPT-4, BARD, and BingAI, and the first responses from each prompt were recorded. These three LLMs were selected as the three most readily available and widely used LLMs in medical research (24). Extreme care was taken to craft each prompt, to ensure that there were no grammatical errors or points of contention. Subsequent clarification of answers or corrections was also not utilised. To preserve the integrity of the original response and mirror the conditions of a junior doctor, the function to regenerate answers or to alter previous responses was not utilised.

All prompts were inputted on the same day on a single account of ChatGPT plus (owned by one of the authors, IS), which provided access to ChatGPT-4. Access to BARD and BingAI required no additional paywall. No inclusion or exclusion criteria were placed on the answers from ChatGPT-4, BARD and BingAI. Institutional ethics were not required as no human participants were involved in the study.

3. Results

3.1. Quantitative Analysis

Through a comprehensive quantitative analysis of ChatGPT, BARD, and BingAI as shown in Table 2, significant variability was observed in the mean readability, as assessed by three standard grading scales. BARD's responses displayed the greatest readability, scoring 50.13 (± 5.00) in the FRE, 9.33 (± 0.76) in the FKGL, and 11.67 (± 0.58) in the CLI.

Regarding the accuracy of the information, ChatGPT surpassed others by providing medical advice closely aligned with current clinical guidelines and up-to-date references. It is evidenced in the highest DISCERN score of 71.7 (± 2.52) compared to BingAI's 64.3 (± 2.08) and BARD's

Table 1. Evaluation of Large Language Model Platforms' Responses

Criteria	ChatGPT	Bing's AI	Google's BARD
The large language model provides accurate answers to questions.;	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [x] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is reliable when generating factual information.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[x] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is proficient at understanding complex questions and providing appropriate answers.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[x] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model provides comprehensive information when answering questions.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model generates content that covers all relevant aspects of a subject.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[x] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is able to provide in-depth information for a wide range of topics.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[x] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is a valuable source of general knowledge.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [x] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is well-versed in a variety of subjects.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [x] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model can provide useful insights and perspectives on various topics.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [x] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model rarely makes errors when referencing sources.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree
The large language model is consistent in providing accurate citations.	[ ] 1 - Strongly Disagree; [ ] 2 - Disagree; [x] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree	[ ] 1 - Strongly Disagree; [x] 2 - Disagree; [ ] 3 - Neither Agree or Disagree; [ ] 4 - Agree; [ ] 5 - Strongly Agree

Table 2. Quantitative Analysis of ChatGPT, BARD, and BingAI

	Readability			Reliability
	FRE	FKGL	CLI	DISCERN
ChatGPT	28.20 ± 1.59	13.80 ± 2.18	14.37 ± 1.10	71.70 ± 2.52
BARD	50.13 ± 5.00	9.33 ± 0.76	11.67 ± 0.58	56.70 ± 2.52
BingAI	22.80 ± 15.16	18.33 ± 4.58	12.67 ± 1.53	64.30 ± 2.08

Abbreviations: FRE, Flesch Reading Ease Score; FKGL, Flesch-Kincaid Grade Level; CLI, Coleman-Liau index; DISCERN, DISCERN questionnaire.

56.7 (± 2.52). A *t*-test comparing all three AIs demonstrated statistical significance in the reliability tests among them, with a P-value < 0.05. Among the readability tests, only the comparison between ChatGPT and BARD was statistically significant (<0.05).

3.2. Qualitative Analysis

Scenario A illustrates a patient who is two-days post-haemorrhoidectomy that begins to deteriorate on the ward (Appendix 1 in the Supplementary File). The guidance that is offered by ChatGPT-4 is to start with a focused history, physical examination, then to review investigations. This dynamic and formulaic process of history, examination, and investigation forms the crux of patient assessment and is a distinguishing characteristic of a master clinician (25). From the onset, there is also general advice to escalate to either a supervising physician or surgeon, which demonstrates an awareness

of limitations and a high level of patient safety. While the scenario is deliberately broad, the complaints of suprapubic pain could have been triggered by a urinary tract infection. The answer may have been improved by further investigation with urine microscopy, culture, and sensitivity.

In our scenario, the patient becomes acutely unwell after spiking a temperature and experiencing mild tachycardia. The recommendation from ChatGPT-4 is to involve a supervising physician or specialist, which is considered safe and appropriate. While the patient has a fever and warrants further investigations, the answer could have been improved by mentioning the importance of conducting a basic septic screen (chest x-ray, urine culture, and wound/blood cultures). As the patient begins to deteriorate and becomes haemodynamically unstable, the situation becomes highly concerning for necrotising fasciitis. While a formal diagnosis is not mentioned, the recommendations for investigations, further imaging, empiric antibiotic therapy, and fluid resuscitation are all appropriate. Additionally, due to the rapidly progressive nature of necrotising fasciitis (26), there is also a recommendation for close monitoring and a low threshold for escalation to intensive care. Once a clinical diagnosis is established, urgent surgical debridement, antibiotic therapy, and fluid resuscitation become the cornerstone of management (27), which are all recommended by ChatGPT-4. The recommendations from this scenario showcase a safe and practical approach to managing a deteriorating surgical patient on the ward.

Scenario B describes a post-operative patient who is tachycardic and hypotensive on the ward (Appendix 2 in the Supplementary File). Again, the recommendation of further history, physical examination, and investigations are used to assess this patient. Early consultation with a supervising physician or surgeon and appropriate treatment with intravenous fluids is also recommended, demonstrating a high level of patient safety. As the patient continues to deteriorate, important differentials of bowel obstruction or post-operative ileus are proposed for consideration, which normally occur 24 - 48 hours post-operatively (28). As the scenario progresses with no recordable drain outputs and persistent pain, ChatGPT-4's response is to reassess the patient, re-evaluate analgesia, and consult with a senior. While the response is adequate and safe, the answer may have been improved by further elucidating options for post-operative analgesia – given its persistent nature (29). Additionally, while there is low evidence for the use of abdominal drainage after open appendectomy (30), the lack of drain outputs may also be misleading, and focus should have been directed toward ascertaining the accuracy of drain measurements.

The progression of the scenario finds a twisted knot in the drain, which likely would have prevented any knowledge of a leak, bleed, or collection. Once a cause is established, the advice to involve the surgical team, fluid resuscitation, and blood transfusion are safe principles of management. In ChatGPT-4's response, there is also the suggestion that this patient may require further surgical re-exploration or haemostasis if ongoing bleeding. This answer therefore may have been improved by ensuring the patient had a valid group and save and was alerted about the possibility of further surgery. Nevertheless, in all answers from ChatGPT-4, safe management principles were complemented with a suggestion to involve a senior medical colleague, highlighting its safe-guarding and practical approach to managing a post-operative patient.

Scenario C describes a challenging patient who has recurrent small bowel obstructions in the emergency department (Appendix 3 in the Supplementary File). The initial approach to diagnosis and management of the patient's condition is congruent with current guidelines (31) – including bowel decompression with nasogastric tube insertion, pain management, and keeping the patient fasted. It is noted that ChatGPT-4 also advises involving the surgical team early in this scenario, demonstrating a strong predisposition towards patient safety. As the patient's electrolytes become deranged, further management of re-checking laboratory values and electrolyte replacement are correctly suggested (32).

The scenario progresses with the patient becoming angry, removing his nasogastric tube, and threatening to discharge him against medical advice. During this ethical dilemma, there is a careful balance between respecting patient wishes and ensuring the best medical practice. The response by ChatGPT-4 highlights the importance of staying calm, educating the patient, offering alternatives, careful documentation, and safety-netting with a discharge plan – highlighting its awareness of patient safety and medico-legal risk. The concept of capacity is also proposed, where junior doctors must ascertain whether a patient fully comprehends the implications of their actions. While the response from ChatGPT-4 was safe and explored issues around capacity, this response may have been improved by listing complications of small bowel obstruction – pain, bowel necrosis and perforation, intra-abdominal abscess, and aspiration (33, 34).

The expert review framework revealed distinct differences in the performance of the three LLMs. ChatGPT-4 outperformed Bing's AI and Google's BARD in providing accurate answers, generating factual information, understanding complex questions, and offering comprehensive, relevant, and in-depth information across various topics. ChatGPT-4 also excelled

in general knowledge and providing useful insights on different subjects. Divergences in the readability and comprehensibility of LLMs are also noted in previous literature (24). It is potentially attributable to varying training data, data pre-processing strategies, and inherent data structures. Such variations could impact each model's proficiency in managing unique terminologies and abbreviations. However, all three models demonstrated room for improvement in referencing sources and providing accurate citations, with none of them scoring above "Neither Agree or Disagree" in those categories.

4. Discussion

This study evaluated and compared the performance of the three most popular LLMs – ChatGPT-4, BingAI, and BARD in providing precise and reliable advice for junior doctors in different post-operative scenarios. Overall, ChatGPT-4 demonstrated a strong foundation in clinical assessment by recommending a structured approach consisting of a focused history, and physical examination, followed by investigations. It consistently recommends junior doctors escalate or involve senior surgeons at an early stage, showcasing an appropriate level of patient safety and awareness of junior doctors' limitations in handling surgical emergencies. ChatGPT-4 was able to generate appropriate differential diagnoses in all 3 scenarios, indicating a comprehensive understanding of the clinical context. It also provided safe and practical guidance on patient management in line with established clinical guidelines while considering the ethical and medico-legal aspects of patient care, including respecting patient autonomy and addressing capacity. However, some ChatGPT-4 responses lacked specificity and comprehensiveness, as shown in scenario B where its answer can be improved by ensuring the patient had a valid group and save for potential re-exploration, reflecting its weakness in anticipating complications.

In contrast, the responses generated by BingAI and BARD are notably less specific in offering distinct recommendations for pathology and imaging tests. Nevertheless, it is important to acknowledge that BARD is the only model to make a preliminary diagnosis of necrotising fasciitis based on haemodynamic instability and provide relevant risk factors to aid junior doctors in formulating differential diagnoses. This capability may indicate BARD's potential as a valuable complementary diagnostic aid (10). Despite these strengths, the performance of BingAI and BARD in providing management advice for handling postoperative emergencies is markedly inferior to that of ChatGPT-4. This deficiency is not only attributable to their responses' lack of structure and comprehensiveness but also to the occasional dissemination of misleading information. For instance, BARD entirely overlooked the inclusion of nasogastric tube (NGT) decompression in the context of bowel obstruction, while BingAI neglected to mention fluid resuscitation in instances of hemodynamic instability. Both oversights could lead to delayed treatment and potentially life-threatening consequences for patients. Nonetheless, both BingAI and BARD consistently emphasised the importance of escalating care throughout their responses. This emphasis is integral to ensuring safe practice for junior doctors in the clinical setting.

The COVID-19 pandemic has substantially disrupted surgical education, posing unique challenges for junior doctors and medical students. The suspension of clinical rotations and postponement of elective surgeries has restricted trainees' clinical exposure. While online learning offers flexibility and convenience, this shift has limited hands-on learning opportunities, especially concerning procedural and physical examination skills. While most universities have resumed face-to-face training, it is also imperative for surgical training programs to consider adapting their curricula for innovative educational strategies, including hybrid learning model and AI-assisted technologies (35). However, the real question posed to medical educators, is how to incorporate this new educational tool without compromising clinical acumen and patient safety.

LLMs particularly have considerable potential to help address the challenges in surgical education due to its interactive nature and rapid information retrieval capacity (36, 37). While they cannot entirely replace in-person training or clinical exposure, they can be instrumental in supplementing and enhancing the educational experience for junior surgeons. Integrating ChatGPT-4 into clinical settings could offer several benefits. Firstly, it can provide real-time assistance to trainees in understanding complex clinical scenarios, interpreting medical data, and making informed decisions. This on-demand support may reduce the cognitive burden on junior doctors and help them refine their clinical reasoning skills. ChatGPT-4 can potentially fill the gaps left by traditional teaching methods, which might be constrained by time, resources, and the availability of experienced faculty. If used as an independent bedside tutor, it can deliver personalised, tailored learning experiences that address the specific needs and knowledge gaps of each trainee. Providing advice on relevant and important questions during history taking and critical components to examination for each presenting complaint. This can facilitate self-directed clinical training and accelerate learning, improving the

overall quality of medical education. AI-generated clinical simulations, although not yet developed, hold great potential as a research area. They could be tailored to simulate various patient scenarios and conditions, offering a safe and controlled learning environment for students to hone their skills.

Although utilising LLMs (especially ChatGPT-4) in assisting surgical education is promising, several ethical concerns and potential challenges must be addressed to ensure the responsible and effective integration of such models into medical training. LLMs can generate information based on patterns in the data they have been trained on, which might not be accurate or up to date as demonstrated in some responses from BARD and BingAI. Relying on potentially incorrect information in medical education can have serious consequences for patient care. Establishing responsibility for the consequences of AI-generated advice is unclear. When a trainee follows a LLM's guidance resulting in negative patient outcomes, determining accountability and liability may also prove challenging. Therefore, effective and ethical use of LLMs in medical education requires adherence to key guidelines. Trainees should view LLMs as advisory tools, verifying their output against other trusted resources since these AI models lack real-world clinical judgement. Data privacy and confidentiality must be prioritised, given that interactions with public LLMs may be stored for future model training, making it essential to avoid sharing personally identifiable or confidential patient information. In addition, there is a risk that surgical trainees may become overly reliant on AI-generated advice, potentially undermining their critical thinking and decision-making abilities (38). Striking a balance between using LLMs as a supplementary tool and developing independent clinical judgment is essential. While AI can provide valuable input, it cannot clinically assess the patient, take a detailed history, or complete an effective examination, all fundamental skills integral to medical training. Thus, supervision by medical professionals is essential, particularly in the early stages, to ensure the accurate interpretation and application of LLM-generated advice, and that surgical trainees continue to develop clinical and communication skills, as well as empathy for their patients.

4.1. Limitations and Future Directions

The primary limitation of this study lies in the fact that the inquiries posed to LLMs are derived from simulated scenarios constructed by a limited group of junior surgical doctors. Consequently, this approach may result in findings that are less generalisable and applicable to a broader context. Nonetheless, the study offers insights into the potential integration of LLMs within surgical education, thereby contributing to the ongoing discourse on artificial intelligence in medical training. Large-scale longitudinal studies should be conducted to continuously assess the impact of this innovative teaching approach on surgical trainees' knowledge, skill development, and overall educational outcomes, with a focus on comparisons with traditional educational methods to identify areas of improvement or potential drawbacks.

Future research on expanding the AI model's training data is also worthwhile to refine the accuracy of their responses. This may be achieved by including more high-quality and up-to-date resources, such as surgical textbooks, guidelines, and research articles specifically covering a wide variety of clinical scenarios, thus providing more accurate and contextually relevant recommendations (39). Collaboration with medical professionals and educators should be encouraged to validate, review, and curate AI-generated content aligning with expert consensus and best medical practices. Other strategies such as the integration of LLMs with existing technologies including virtual reality and surgical simulators, in the future, may further enhance the learning experience. Additionally, amalgamating LLMs into virtual and robotic trainers may also provide more comprehensive and context-aware guidance to the surgical trainee, leading to improved delivery of feedback, and ultimately improving patient outcomes (40).

4.2. Conclusions

This study illustrates the potential of using AI technologies to aid junior doctors by providing accurate and pertinent guidance in common ward-based surgical scenarios. The findings suggest LLMs, particularly ChatGPT-4, hold promise as valuable educational resources in medical training in certain scenarios. However, while these results are promising, ethical concerns and challenges limit the routine use of LLM in medical education. Further investigations are warranted to examine the applicability of LLM in diverse medical specialties, as well as its impact on patient outcomes and building clinical acumen in junior doctors. By comprehending the advantages and constraints of AI language models in medical education, we may devise innovative approaches to instructing future generations of healthcare professionals.

Supplementary Material

Supplementary material(s) is available here [To read supplementary materials, please refer to the journal website and open PDF/HTML].

Acknowledgments

We would like to acknowledge the research staff of Dr A.L.'s research group and Surgical Outcomes Research Centre (SOuRCe).

Footnotes

Authors' Contribution: AHYS: Study concept and design, acquisition/analysis/interpretation of data, drafted the manuscript, critically revised the manuscript. DG: Student concept and design, acquisition/analysis/interpretation of data, drafted the manuscript, critically revised the manuscript. XM: Drafted the manuscript, critically revised the manuscript. IS: Drafted the manuscript, critically revised the manuscript. ACWS: Drafted the manuscript, critically revised the manuscript. DD: Study supervision, analysis of data, drafted the manuscript, critically revised the manuscript. AL: Study supervision, analysis of data, drafted the manuscript, critically revised the manuscript.

Conflict of Interests: AHYS and ACWS would like to declare they are brothers. However, both have contributed significantly to the work of this paper. AHYS would like to declare that he is a member of the AMA (NSW).

Data Reproducibility: The dataset presented in the study is available on request from the corresponding author during submission or after publication.

Ethical Approval: Institutional ethics were not required as no human participants, people, medical records, or animal studies were involved in this study.

Funding/Support: There was no funding/support for this manuscript.

References

1. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;**33**:1877–901.
2. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *npj Dig Med*. 2022;**6**(75 (2023)). https://doi.org/10.1038/s41746-023-00819-6.
3. Hacker P, Engel A, Mauer M. Regulating ChatGPT and other Large Generative AI Models. *arXiv preprint arXiv:230202337*. 2023:1112–23. https://doi.org/10.1145/3593013.3594067.
4. Seth I, Sinkjaer Kenney P, Bulloch G, Hunter-Smith DJ, Bo Thomsen J, Rozen WM. Artificial or Augmented Authorship? A Conversation with a Chatbot on Base of Thumb Arthritis. *Plast Reconstr Surg Glob Open*. 2023;**11**(5). e4999. [PubMed ID: 37250832]. [PubMed Central ID: PMC10219695]. https://doi.org/10.1097/GOX.0000000000004999.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;**2**(2). e0000198. [PubMed ID: 36812645]. [PubMed Central ID: PMC9931230]. https://doi.org/10.1371/journal.pdig.0000198.
6. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2023. [PubMed ID: 36916887]. https://doi.org/10.1002/ase.2270.
7. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. *Front Artif Intell*. 2023;**6**:1166014. [PubMed ID: 37091303]. [PubMed Central ID: PMC10113434]. https://doi.org/10.3389/frai.2023.1166014.
8. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. *Neurosurgery*. 2023. [PubMed ID: 37581444]. https://doi.org/10.1227/neu.0000000000002632.
9. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved Performance of ChatGPT-4 on the OKAP Examination: A Comparative Study with ChatGPT-3.5. *J Acad Ophthalmol (2017)*. 2023;**15**(2):e184–7. [PubMed ID: 37701862]. [PubMed Central ID: PMC10495224]. https://doi.org/10.1055/s-0043-1774399.
10. Seth I, Bulloch G, Rozen WM. Applications of Artificial Intelligence and Large Language Models to Plastic Surgery Research. *Aesthet Surg J*. 2023;**43**(10):NP809–10. [PubMed ID: 37392428]. https://doi.org/10.1093/asj/sjad210.
11. Smith JA, Kaye AH, Christophi C, Brown WA. *Textbook of surgery*. John Wiley & Sons; 2020.
12. Farne H, Norris-Cervetto E, Warbrick-Smith J. *Oxford cases in medicine and surgery*. Oxford University Press; 2015.
13. Nemoto T, Beglar D. Developing Likert-scale questionnaires.In: Sonda N, Krause A, editors. *JALT 2013 conference proceedings*. Tokyo: JALT. 2014. p. 1–8.
14. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. 1999;**53**(2):105–11. [PubMed ID: 10396471]. [PubMed Central ID: PMC1756830]. https://doi.org/10.1136/jech.53.2.105.
15. England GW, Thomas M, Paterson DG. Reliability of the original and the simplified Flesch reading ease formulas. *J Appl Psychol*. 1953;**37**(2):111–3. https://doi.org/10.1037/h0055346.
16. Flesch R. Flesch-Kincaid readability test. *Retrieved October*. 2007;**26**(3):2007.
17. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;**60**(2):283–4. https://doi.org/10.1037/h0076540.
18. Stake RE. *The art of case study research*. SAGE; 1995.
19. Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: a tool for assessing the quality of written patient information on treatment choices. *Patient Educ Couns*. 2002;**47**(3):273–5. [PubMed ID: 12088606]. https://doi.org/10.1016/s0738-3991(01)00225-7.
20. Cassidy JT, Baker JF. Orthopaedic Patient Information on the World Wide Web: An Essential Review. *J Bone Joint Surg Am*. 2016;**98**(4):325–38. [PubMed ID: 26888683]. https://doi.org/10.2106/JBJS.N.01189.
21. Kher A, Johnson S, Griffith R. Readability Assessment of Online Patient Education Material on Congestive Heart Failure. *Adv Prev Med*. 2017;**2017**:9780317. [PubMed ID: 28656111]. [PubMed Central ID: PMC5471568]. https://doi.org/10.1155/2017/9780317.
22. Szmuda T, Ozdemir C, Ali S, Singh A, Syed MT, Sloniewski P. Readability of online patient education material for the novel coronavirus disease (COVID-19): a cross-sectional health literacy study. *Public Health*. 2020;**185**:21–5. [PubMed ID: 32516624]. [PubMed Central ID: PMC7260546]. https://doi.org/10.1016/j.puhe.2020.05.041.
23. Wadhwa RR, Marappa-Ganeshan R. Disclosure: Raghavendra Marappa-Ganeshan declares no relevant financial relationships with ineligible companies. *T Test*. Treasure Island, USA: StatPearls Publishing; 2023.
24. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman HP. Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. *medRxiv*. 2023. https://doi.org/10.1101/2023.06.04.23290786.

25. Davis JL, Murray JF. History and physical examination. *Murray and Nadel's Textbook of Respiratory Medicine*. **263**. Elsevier; 2016.
26. Randhawa S, Bhullar JS, Rana G, Bhullar A, Mittal VK, Goriel Y. Necrotizing fasciitis–a sinister complication of hemorrhoidectomy. *Int J Colorectal Dis*. 2015;**30**(6):851–2. [PubMed ID: 25367181]. https://doi.org/10.1007/s00384-014-2050-4.
27. Misiakos EP, Bagias G, Papadopoulos I, Danias N, Patapis P, Machairas N, et al. Early Diagnosis and Surgical Treatment for Necrotizing Fasciitis: A Multicenter Study. *Front Surg*. 2017;**4**:5. [PubMed ID: 28224127]. [PubMed Central ID: PMC5293831]. https://doi.org/10.3389/fsurg.2017.00005.
28. Luckey A, Livingston E, Tache Y. Mechanisms and treatment of postoperative ileus. *Arch Surg*. 2003;**138**(2):206–14. [PubMed ID: 12578422]. https://doi.org/10.1001/archsurg.138.2.206.
29. Small C, Laycock H. Acute postoperative pain management. *Br J Surg*. 2020;**107**(2):e70–80. [PubMed ID: 31903595]. https://doi.org/10.1002/bjs.11477.
30. Li Z, Li Z, Zhao L, Cheng Y, Cheng N, Deng Y. Abdominal drainage to prevent intra-peritoneal abscess after appendectomy for complicated appendicitis. *Cochrane Database Syst Rev*. 2021;**8**(8). CD010168. [PubMed ID: 34402522]. [PubMed Central ID: PMC8407456]. https://doi.org/10.1002/14651858.CD010168.pub4.
31. Bower KL, Lollar DI, Williams SL, Adkins FC, Luyimbazi DT, Bower CE. Small Bowel Obstruction. *Surg Clin North Am*. 2018;**98**(5):945–71. [PubMed ID: 30243455]. https://doi.org/10.1016/j.suc.2018.05.007.
32. Tong JWV, Lingam P, Shelat VG. Adhesive small bowel obstruction - an update. *Acute Med Surg*. 2020;**7**(1). e587. [PubMed ID: 33173587]. [PubMed Central ID: PMC7642618]. https://doi.org/10.1002/ams2.587.
33. Naidu K. Small bowel obstruction. *Emergency Surgery for Low Resource Regions*. Springer; 2021. p. 135–40.
34. Schick MA, Kashyap S, Meseeha M. *Small Bowel Obstruction*. Treasure Island, USA: StatPearls Publishing; 2017.
35. Wartman SA, Combs CD. Reimagining Medical Education in the Age of AI. *AMA J Ethics*. 2019;**21**(2):E146–52. [PubMed ID: 30794124]. https://doi.org/10.1001/amajethics.2019.146.
36. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, et al. Evaluating Chatbot Efficacy for Answering Frequently Asked Questions in Plastic Surgery: A ChatGPT Case Study Focused on Breast Augmentation. *Aesthet Surg J*. 2023;**43**(10):1126–35. [PubMed ID: 37158147]. https://doi.org/10.1093/asj/sjad140.
37. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. *Aesthetic Plast Surg*. 2023. [PubMed ID: 37095384]. https://doi.org/10.1007/s00266-023-03338-7.
38. Farrokhnia M, Banihashem SK, Noroozi O, Wals A. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innov Educ Teach Int*. 2023:1–15. https://doi.org/10.1080/14703297.2023.2195846.
39. Sapci AH, Sapci HA. Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review. *JMIR Med Educ*. 2020;**6**(1). e19285. [PubMed ID: 32602844]. [PubMed Central ID: PMC7367541]. https://doi.org/10.2196/19285.
40. Winkler-Schwartz A, Bissonnette V, Mirchi N, Ponnudurai N, Yilmaz R, Ledwos N, et al. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J Surg Educ*. 2019;**76**(6):1681–90. [PubMed ID: 31202633]. https://doi.org/10.1016/j.jsurg.2019.05.015.