ACFS
ARCHIVES OF CRANIOFACIAL SURGERY
Arch Craniofac Surg Vol.25 No.3, 116-122
https://doi.org/10.7181/acfs.2024.00115

The new frontier: utilizing ChatGPT to expand craniofacial research

Andi Zhang[1], Ethan Dimock[2], Rohun Gupta[1], Kevin Chen[1]

[1]*Division of Plastic and Reconstructive Surgery, Saint Louis University School of Medicine, St. Louis, MO, USA*
[2]*Oakland University William Beaumont School of Medicine, Rochester, MI, USA*

Background: Due to the importance of evidence-based research in plastic surgery, the authors of this study aimed to assess the accuracy of ChatGPT in generating novel systematic review ideas within the field of craniofacial surgery.
Methods: ChatGPT was prompted to generate 20 novel systematic review ideas for 10 different subcategories within the field of craniofacial surgery. For each topic, the chatbot was told to give 10 "general" and 10 "specific" ideas that were related to the concept. In order to determine the accuracy of ChatGPT, a literature review was conducted using PubMed, CINAHL, Embase, and Cochrane.
Results: In total, 200 total systematic review research ideas were generated by ChatGPT. We found that the algorithm had an overall 57.5% accuracy at identifying novel systematic review ideas. ChatGPT was found to be 39% accurate for general topics and 76% accurate for specific topics.
Conclusion: Craniofacial surgeons should use ChatGPT as a tool. We found that ChatGPT provided more precise answers with specific research questions than with general questions and helped narrow down the search scope, leading to a more relevant and accurate response. Beyond research purposes, ChatGPT can augment patient consultations, improve healthcare equity, and assist in clinical decision-making. With rapid advancements in artificial intelligence (AI), it is important for plastic surgeons to consider using AI in their clinical practice to improve patient-centered outcomes.

Abbreviations: AI, artificial intelligence; LLM, large language model; OKAP, Ophthalmic Knowledge Assessment Program; USMLE, United States Medical Licensing Exam

Keywords: Artificial intelligence / Craniofacial research / Machine learning / Natural language processing / Systematic review

Original Article

INTRODUCTION

Recent advancements in large language model (LLM) technology have transformed the field of artificial intelligence (AI). A prime example of this is ChatGPT (Chat Generative Pretrained Transformer), a LLM released to the public in November 2022 by OpenAI, a company based in San Francisco, USA. LLMs are pretrained transformer models that are self-supervised and can be adapted with fine-tuning to a wide range of natural language tasks [1]. ChatGPT is part of a line of LLMs that achieve the state-of-the-art performance with minimal prerequisite fine-tuning [1]. Specifically, ChatGPT has been trained on a wide variety of conversational prompts and webpages to encourage dialogue output that fosters interactions in a more humanistic fashion, allowing for seemingly natural human-like conversations. In short, ChatGPT is one of the most sophisticated and widely available natural language AI models.

ChatGPT has been quickly put to the test in the medical field

Correspondence: Andi Zhang
Division of Plastic and Reconstructive Surgery, Saint Louis University School of Medicine, SLUCare Academic Pavilion 1008 S. Spring Ave, Suite 1500 St. Louis, MO 63110, USA
E-mail: andyzhang214@gmail.com

How to cite this article:
Zhang A, Dimock E, Gupta R, Chen K. The new frontier: utilizing ChatGPT to expand craniofacial research. Arch Craniofac Surg 2024;25(3):116-122. https://doi.org/10.7181/acfs.2024.00115

Received February 29, 2024 / Revised April 5, 2024 / Accepted June 11, 2024

www.e-acfs.org
pISSN 2287-1152
eISSN 2287-5603

and extensively evaluated. Previous studies have demonstrated that ChatGPT successfully passed all three sections of the United States Medical Licensing Exam (USMLE) [2]. Other studies tested ChatGPT on more advanced exams, such as the plastic surgery in-service exam, Ophthalmic Knowledge Assessment Program (OKAP) exam in ophthalmology, and general surgery board exams. ChatGPT was able to achieve human-level performance across all exams: 60.2% for the USMLE (60.2%), 61% for the OKAP, 57% for the Plastic Surgery In-service Training Examination, and 76.4% for the General Surgery Board Examination [2-5]. Despite the impressive results generated by ChatGPT on standardized exam questions, its performance in complex real-world scenarios, particularly in demanding fields such as medicine that involve high cognitive loads, remains uncertain [6]. ChatGPT represents a new line of AI models that combine vast clinical knowledge with novel conversational abilities. ChatGPT's ability to produce narrative-like answers has a wide variety of use cases in research, and many authors have used its writing ability to write papers or summarize results; however, ChatGPT's ability to generate novel research ideas has not yet been widely tested [7].

Plastic and reconstructive surgery has significantly evolved over the past several decades. Innovations in technology, improved understanding of disease pathogenesis, and superior interventions have led to improved patient outcomes and propelled constant progress in clinical and basic science research [8,9]. Craniofacial surgery in particular, requires long patient follow-up to observe results and relies heavily on past research to influence the paradigms of the present and future. Systematic reviews in craniofacial surgery help guide treatment decisions; however, generating novel systematic reviews requires a significant investment of time in reviewing the literature. A recent study by Gupta et al. [10] presented evidence that ChatGPT can be utilized to produce novel systematic review ideas related to cosmetic plastic surgery. This study sought to determine whether ChatGPT may also be applied to generate novel systematic review ideas to reflect the diverse practices—both operative and nonoperative—in the field of craniofacial plastic surgery.

METHODS

ChatGPT was prompted to generate 20 novel systematic review ideas for 10 different subcategories within the field of craniofacial surgery on January 15, 2023. For each subcategory, 10 of these 20 responses were chosen to be "non-specific," while the other 10 were split evenly between two specific areas within that subcategory. These responses were significantly narrower in scope and hence categorized as "specific." For instance, ChatGPT was given the following three separate commands: (1) "Give me 10 novel systematic review ideas that have not been published regarding cleft lip." (Table 1, Fig. 1); (2) "Give me 5 novel systematic review ideas that have not been published related to lip adhesion."; or (3) "Give me 5 novel systematic review ideas that have not been published related to Millard repair."

Table 1. Ten ChatGPT-generated systematic review topics relating to cleft lip, the corresponding number of non-systematic and systematic reviews already published on the topic, and novelty status

Research topic	No. of non-systematic reviews published	No. of systematic reviews published	Novel idea?
A systematic review of the long-term outcomes of cleft lip repair surgery in children, including patient satisfaction, functional outcomes, and quality of life.	2,449	54	N
A systematic review of the effectiveness of different surgical techniques for cleft lip repair, including comparative studies of outcomes and complications.	56	5	N
A systematic review of the psychological impact of cleft lip on patients and their families, including the incidence of depression, anxiety, and other mental health issues.	615	17	N
A systematic review of the economic burden of cleft lip repair surgery, including the cost of surgery, follow-up care, and potential long-term healthcare costs.	198	1	N
A systematic review of the impact of cleft lip on speech development in children, including the incidence of speech disorders and the effectiveness of speech therapy.	510	2	N
A systematic review of the impact of cleft lip on dental development in children, including the incidence of dental anomalies and the effectiveness of orthodontic treatment.	227	1	N
A systematic review of the genetic and environmental factors that contribute to the development of cleft lip, including risk factors and potential prevention strategies.	293	4	N
A systematic review of the incidence and prevalence of cleft lip in different populations and regions around the world, including differences in gender, ethnicity, and socioeconomic status.	106	0	Y
A systematic review of the effectiveness of multidisciplinary cleft lip teams in improving patient outcomes and reducing complications.	57	0	Y
A systematic review of the impact of cleft lip on social functioning and relationships, including the incidence of social stigmatization and discrimination.	24	1	N

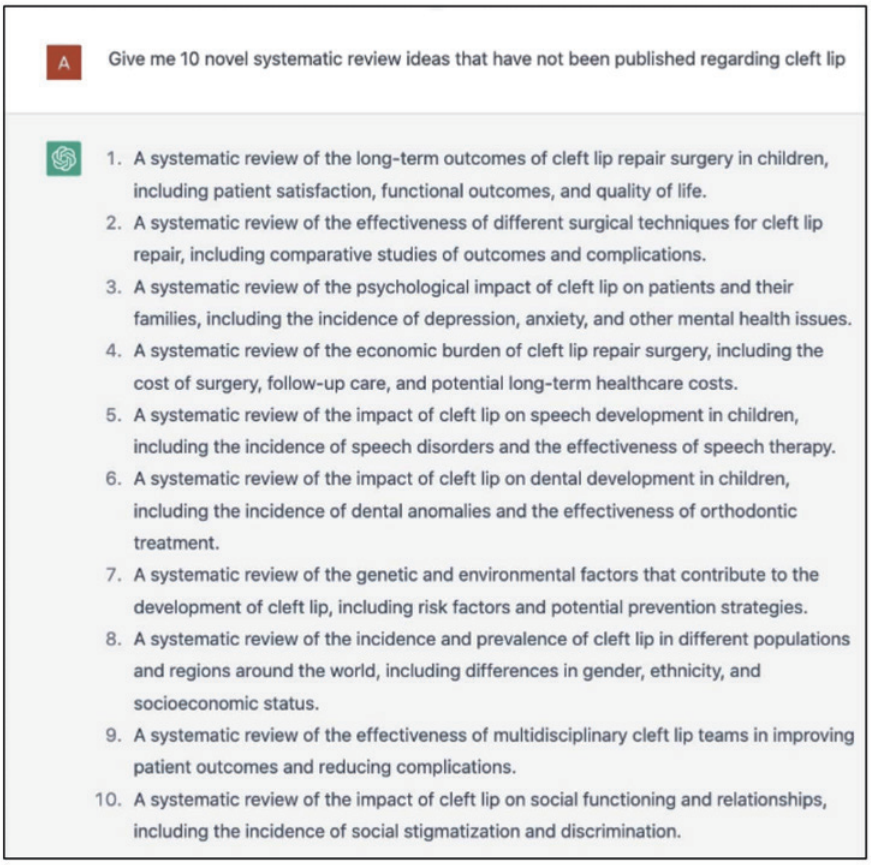

Fig. 1. The ChatGPT-generated novel systematic review ideas that have not been published regarding cleft lip.

Table 2. Five ChatGPT-generated systematic review topics relating to lip adhesion, the corresponding number of non-systematic and systematic reviews already published on the topic, and novelty status

Research topic	No. of non-systematic reviews published	No. of systematic reviews published	Novel idea?
A systematic review of the incidence and risk factors for recurrent lip adhesions after surgical treatment, including the effectiveness of preventive measures.	6	0	Y
A systematic review of the impact of lip adhesions on social functioning and relationships, including the incidence of social stigmatization and discrimination.	3	0	Y
A systematic review of the economic burden of lip adhesions after cleft lip repair surgery, including the cost of treatment, follow-up care, and potential long-term healthcare costs.	3	0	Y
A systematic review of the impact of lip adhesions on speech development in children with cleft lip, including the incidence of speech disorders and the effectiveness of speech therapy.	6	0	Y
A systematic review of the impact of lip adhesions on dental development in children with cleft lip, including the incidence of dental anomalies and the effectiveness of orthodontic treatment.	21	0	Y

A "novel idea" was operationally defined as a research topic that ChatGPT generated that was both medically accurate and had no previous systematic reviews published. An idea was considered medically accurate when published research existed on the topic from a literature search. In order to assess the accuracy of the responses synthesized by ChatGPT, a literature search was conducted using PubMed (National Institutes of Health), CINAHL (EBSCO Information Services), Embase (Elsevier), and Cochrane (Cochrane Library) to cover the general literature and determine the number of systematic reviews that had been published on each topic.

The data were collected and subsequently analyzed by two reviewers independently for concordance on Google Sheets. Any disagreements were discussed to reach consensus. Formulated tables were created to complete a statistical analysis and interpret various aspects of the data. For example, three primary variables were assessed for each generated idea: the number of non-systematic reviews published, the number of systematic

Table 3. Five ChatGPT-generated systematic review topics relating to Millard repair, corresponding number of non-systematic and systematic reviews already published on topic, and novelty status

Research topic	No. of non-systematic reviews published	No. of systematic reviews published	Novel idea?
A systematic review of the impact of Millard repair on speech development in children with cleft lip, including the incidence of speech disorders and the effectiveness of speech therapy.	3	0	Y
A systematic review of the impact of Millard repair on dental development in children with cleft lip, including the incidence of dental anomalies and the effectiveness of orthodontic treatment.	9	0	Y
A systematic review of the psychological impact of Millard repair on patients and their families, including the incidence of depression, anxiety, and other mental health issues.	2	0	Y
A systematic review of the economic burden of Millard repair for cleft lip, including the cost of surgery, follow-up care, and potential long- term healthcare costs.	1	0	Y
A systematic review of the effectiveness of different anesthetic techniques for Millard repair of cleft lip, including comparative studies of outcomes and complications.	2	0	Y

reviews published, and whether the idea was novel (Tables 1-3). Averages of the three primary variables were also calculated for each subcategory topic, which is a simple average of the subcategory topic's 20 generated ideas. An idea was considered novel if the topic was dealt with by non-systematic review publications, but there were no related systematic reviews.

RESULTS

In total, ChatGPT generated 200 research ideas for 10 unique subcategory topics spanning the field of craniofacial surgery. Overall, we found that ChatGPT was able to create 115 novel systematic review ideas, with a 57.5% accuracy rate. When stratified by general and specific topics, it was determined that ChatGPT was 39% accurate for general craniofacial surgery topics and 76% accurate for specific topics (Table 4).

The subcategory topics garnering the highest average number of non-systematic reviews were vascular malformations (n = 700), cleft palate (n = 635), cleft lip (n = 547), and facial trauma (n = 663). Notably, these subcategories also had the highest average number of systematic reviews published, with 13.2 for vascular malformations, 13.1 for cleft palate, 10.6 for cleft lip, and 9.0 for facial trauma. These statistics inversely correlate with the generation of novel research ideas in the "general" category for these subcategories. Topics with a greater number of published studies yielded fewer novel systematic review ideas.

Table 4. Novel ChatGPT-generated systematic review topics as a percentage of the overall topics generated for each category

Research topic	No. of novel general ideas (% out of 100)	No. of novel specific ideas (% out of 100)
Cleft lip	20	100
Cleft palate	20	70
Alveolar cleft	70	60
Craniosynostosis	60	90
Orthognathic surgery	40	70
Pierre robin sequence	50	50
Distraction osteogenesis	30	60
Vascular malformations	40	100
Facial trauma	10	70
Microtia	50	90
Average	39	76

DISCUSSION

Our study revealed that ChatGPT has the potential to assist clinicians and researchers in generating novel ideas for systematic reviews on a variety of topics within the field of craniofacial plastic surgery. The overall accuracy of 57.5% achieved by ChatGPT in this study is consistent with previous research by Gupta et al. [11], which reported an overall accuracy of 55% in generating novel systematic review ideas for aesthetic plastic surgery. When stratified into general and specific topics, we found that ChatGPT was able to achieve a higher accuracy of 76% in generating specific topics versus an accuracy of 39% in general topics. Thus, we propose that users of ChatGPT should utilize questions and prompts that are more specific to their research topic to generate more accurate answers. With its generative power and accuracy in creating novel systematic review ideas, ChatGPT can assist the research process in the following ways. First, ChatGPT can be used as a tool to generate new and innovative ideas for systematic reviews in the field of craniofacial surgery. Providing more specific prompts and questions related to craniofacial plastic surgery allows ChatGPT to be more accurate with its answers and topic generation. Researchers and cli-

nicians can interact with ChatGPT to explore various aspects of craniofacial surgery and generate unique research ideas. Second, by utilizing specific prompts, researchers can employ ChatGPT to explore specific topics within craniofacial surgery. ChatGPT can help identify subtopics, potential research questions, and relevant areas that may warrant further investigation. Third, ChatGPT can provide a quick and efficient way to explore research ideas in the field of craniofacial surgery. Researchers can interact with the system, ask questions, and receive prompt responses, allowing for a more streamlined and time-saving process when compared to more traditional and manual research methods.

The observation that ChatGPT becomes more accurate with more specific prompts has been previously reported and is an area of active research known as "prompt engineering" [12,13]. While a prompt may range from just a few words to a few pages long, its goal is to direct the LLM to provide the desired content or output. Regardless of the specific task at hand, general best practices for prompt engineering include: (1) being precise with specific wording and avoiding broad/open-ended prompts; (2) providing examples when applicable as if you were telling a human; or (3) utilizing interactive refinements, such as asking, "do you understand?" and providing clarification, validation, or redirection. Applying these techniques enables the user to maximize ChatGPT's abilities and increases the quality and accuracy of its responses [14].

It is important to note that ChatGPT should be used as a supportive tool in the research process and not as a replacement for critical analysis, human expertise, and peer review. Researchers should always carefully evaluate and validate the generated ideas and information obtained from ChatGPT.

Overall, ChatGPT has the potential to help in the process of creating systematic reviews. By employing more specific questions, users can utilize ChatGPT to generate more accurate research ideas, explore existing topics, and help narrow down the scope of search on a certain topic of interest. This can help researchers lead a more relevant and efficient search.

ChatGPT can also be a useful tool beyond research purposes. Craniofacial surgeons aiming to improve their clinical practice and improve patient experiences may consider the following applications of ChatGPT. First, craniofacial surgeons can use ChatGPT to provide medical information and answer questions tailored to a potential patient during a remote or virtual consultation. Patients can ask questions to learn about certain conditions and procedures such as the recovery process, risks and benefits, and costs. ChatGPT has built-in memory and thus can maintain longitudinal conversations with each patient, so the conversation does not have to start from the beginning every time. Second, ChatGPT can also improve healthcare equity and reduce barriers to care by delivering medical advice and consultations to patients regardless of their location or financial status. This reduces the burden of time and travel for both clinicians and patients and ensures a level of healthcare equity, especially for those in underserved and remote areas. Third, ChatGPT can help optimize clinical decision support alert systems and reduce physician alert fatigue by ensuring that alerts are relevant, justifiable, and timely. AI-powered LLMs like ChatGPT can help clinicians make more informed clinical decisions and reduce the risk of medical errors. Through reinforcement learning from human feedback, ChatGPT can streamline more complex medical decision systems that craniofacial surgeons often use and become more "intelligent" with more experience [15].

Our findings are not without limitations. Although ChatGPT has been trained on several major datasets, its knowledge has the cutoff year of 2021. As a result, some systematic reviews that had been published during 2022 and 2023 were reported as "novel" by ChatGPT. Thus, it is possible that ChatGPT may not present users with the most reliable and up-to-date information. Future updates or installations of ChatGPT have ameliorated this problem, with the latest release of GPT-4o in May 2024, which has a training cutoff date of May 2023. However, unlike its predecessor ChatGPT, GPT-4o also has the added ability to search the web for up-to-date answers beyond its training database [16]. This latest free version, ChatGPT-4o, has since replaced the original free version of ChatGPT tested in this study. Additionally, our study is limited by our sample size and breadth of topics, consisting of only 20 "novel" systematic review ideas for each of the 10 subcategories within the expansive field of craniofacial surgery. Previous literature has described "nonsensical ideas" as topics for which there were no publications because the idea either does not exist or is not a common medical practice [17]. These anomalies were not found in this study. Lastly, because ChatGPT was trained using 8 million web pages, it has the potential to further exacerbate misinformation that already exists and amplify certain hidden biases.

Since its release in 2022, ChatGPT has become popular within the medical community. Researchers have used it to demonstrate its capabilities in many areas, such as taking medical exams, writing research papers, and providing patient education [2,7,11,18,19]. However, it has also been met with concerns regarding its ability to replace or displace traditional physician roles that are centuries old. With the advent of ChatGPT, the way we interact with each other will change, and these changes will affect the medical field in one way or another. Therefore, it is our job as medical professionals to ensure that the technology advances in a way that complements established medical roles

instead of creating conflict. We believe that close collaboration among computer scientists, medical professionals, ethicists, and patients is needed to create AI technologies that will benefit patients and support established clinical workflows. It is important to consider both the patient and the physician perspectives when designing these products, such as balancing patient satisfaction and safety with physician efficiency and well-being. A regulatory framework should be established in the near future to represent the legal and ethical implications of such AI from a medical standpoint and to advocate, review, or voice concerns regarding the uses of generative AI in medicine. As with most applications of AI in medicine, the role of ChatGPT should remain adjunctive to that of the physician and researcher. "Novel" ideas generated by ChatGPT may not necessarily be clinically useful, and these ideas should be used under the guidance of an experienced researcher. Craniofacial surgery is a complex field and ChatGPT can be a useful tool. However, patients in need of craniofacial procedures should always consult qualified surgeons and healthcare professionals.

ChatGPT and AI in general, are becoming powerful research tools. We found that ChatGPT is capable of generating novel systematic review ideas in the field of craniofacial surgery and performs with higher accuracy with more specific research questions, which can help narrow down the initial search scope. This helps researchers conduct a more relevant and efficient search. However, novel ideas may not necessarily translate to useful studies; therefore, while ChatGPT can be helpful in the generation of ideas, the judgment of an experienced researcher is still required. Beyond research purposes, ChatGPT can augment patient consultations, improve healthcare equity, and assist in clinical decision-making. With rapid advancements in AI, it is important for plastic surgeons to consider the use of AI throughout both their clinical and scholarly endeavors.

NOTES

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Funding

None.

ORCID

Andi Zhang https://orcid.org/0009-0005-2851-0736
Ethan Dimock https://orcid.org/0009-0001-8423-3965
Rohun Gupta https://orcid.org/0000-0003-1491-2441
Kevin Chen https://orcid.org/0000-0003-2545-6381

Author contributions

Conceptualization: Andi Zhang, Rohun Gupta, Kevin Chen. Data curation: Andi Zhang, Ethan Dimock. Formal analysis: Andi Zhang. Methodology; Project administration: Andi Zhang, Ethan Dimock, Rohun Gupta, Kevin Chen. Visualization: Andi Zhang, Ethan Dimock, Kevin Chen. Writing – original draft: Andi Zhang, Ethan Dimock. Writing - review & editing: Andi Zhang, Rohun Gupta, Kevin Chen. Investigation: Andi Zhang, Rohun Gupta, Kevin Chen. Resources: Andi Zhang. Supervision: Rohun Gupta, Kevin Chen.

REFERENCES

1. Sejnowski TJ. Large language models and the reverse Turing test. Neural Comput 2023;35:309-42.
2. 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:e0000198.
3. 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:e184-7.
4. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service examination. Aesthet Surg J 2023;43:NP1085-9.
5. Kang E, Gillespie BM, Tobiano G, Chaboyer W. Discharge education delivered to general surgical patients in their management of recovery post discharge: a systematic mixed studies review. Int J Nurs Stud 2018;87:1-13.
6. Kumar B. GPT-1, GPT-2 and GPT-3 models explained: learn the evolution of AI language models [Internet]. 360DigiTMG; c2023 [cited 2023 Sep 29]. https://360digitmg.com/blog/types-of-gpt-in-artificial-intelligence
7. Osmanovic-Thunstrom A, Steingrimsson S, Thunstrom AO. Can GPT-3 write an academic paper on itself, with minimal human input? Archive ouverte HAL [Preprint] 2022 Jun 21 [cite 2023 Sep 29]. https://hal.science/hal-03701250
8. Hansdorfer MA, Horen SR, Alba BE, Akin JN, Dorafshar AH, Becerra AZ. The 100 most-disruptive articles in plastic and reconstructive surgery and sub-specialties (1954-2014). Plast Reconstr Surg Glob Open 2021;9:e3446.
9. Grunwald T, Krummel T, Sherman R. Advanced technologies in plastic surgery: how new innovations can improve our training and practice. Plast Reconstr Surg 2004;114:1556-67.
10. Gupta R, Park JB, Bisht C, Herzog I, Weisberger J, Chao J, et al. Expanding cosmetic plastic surgery research with ChatGPT.

Aesthet Surg J 2023;43:930-7.
11. Gupta R, Herzog I, Park JB, Weisberger J, Firouzbakht P, Ocon V, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023;43:NP1078-82.
12. Gan C, Mori T. Sensitivity and robustness of large language models to prompt template in Japanese text classification tasks. arXiv [Preprint] 2023 Jun 8 [cited 2023 Sep 29]. https://doi.org/10.48550/arXiv.2305.08714
13. Dash D, Thapa R, Banda JM, Swaminathan A, Cheatham M, Kashyap M, et al. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv [Preprint] 2023 May 1 [cited 2023 Sep 29]. https://doi.org/10.48550/arXiv.2304.13714
14. Bhatti BM. The art and science of crafting effective prompts for LLMs [Internet]. Medium; c2023 [cited 2023 Sep 29]. https://thebabar.medium.com/the-art-and-science-of-crafting-effective-prompts-for-llms-e04447e8f96a
15. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30:1237-45.
16. Barr K. GPT-4 is a giant black box and its training data remains a mystery [Internet]. Gizmodo; c2023 [cited 2023 Sep 29]. https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989
17. Gupta R, Herzog I, Najafali D, Firouzbakht P, Weisberger J, Mailey BA. Application of GPT-4 in cosmetic plastic surgery: does updated mean better? Aesthet Surg J 2023;43:NP666-9.
18. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health 2023;13:01003.
19. 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;43:1126-35.