ELSEVIER

SURGERY FOR OBESITY AND RELATED DISEASES

Surgery for Obesity and Related Diseases 20 (2024) 603–608

ASMBS Guidelines/Statements

Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in generating clinician-level bariatric surgery recommendations

Yung Lee, M.D., M.PH.[a,b], Thomas Shin, M.D., Ph.D.[c], Léa Tessier, M.D.[a], Arshia Javidan, M.D., M.Sc.[d], James Jung, M.D., Ph.D.[e], Dennis Hong, M.D., M.Sc.[a], Andrew T. Strong, M.D.[f], Tyler McKechnie, M.D., M.Sc.[a], Sarah Malone, B.H.Sc.[a], David Jin, B.H.Sc.[a], Matthew Kroh, M.D.[f], Jerry T. Dang, M.D., Ph.D.[f,*], ASMBS Artificial Intelligence and Digital Surgery Task Force

[a]*Division of General Surgery, McMaster University, Hamilton, Ontario, Canada*
[b]*Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts*
[c]*Department of Surgery, Brigham and Women's Hospital, Boston, Massachusetts*
[d]*Division of Vascular Surgery, University of Toronto, Toronto, Ontario, Canada*
[e]*Division of General Surgery, Duke University, Durham, North Carolina*
[f]*Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio*

Received 3 March 2024; accepted 9 March 2024

Abstract

Background: The formulation of clinical recommendations pertaining to bariatric surgery is essential in guiding healthcare professionals. However, the extensive and continuously evolving body of literature in bariatric surgery presents considerable challenge for staying abreast of latest developments and efficient information acquisition. Artificial intelligence (AI) has the potential to streamline access to the salient points of clinical recommendations in bariatric surgery.

Objectives: The study aims to appraise the quality and readability of AI-chat-generated answers to frequently asked clinical inquiries in the field of bariatric and metabolic surgery.

Setting: Remote.

Methods: Question prompts inputted into AI large language models (LLMs) and were created based on pre-existing clinical practice guidelines regarding bariatric and metabolic surgery. The prompts were queried into 3 LLMs: OpenAI ChatGPT-4, Microsoft Bing, and Google Bard. The responses from each LLM were entered into a spreadsheet for randomized and blinded duplicate review. Accredited bariatric surgeons in North America independently assessed appropriateness of each recommendation using a 5-point Likert scale. Scores of 4 and 5 were deemed appropriate, while scores of 1–3 indicated lack of appropriateness. A Flesch Reading Ease (FRE) score was calculated to assess the readability of responses generated by each LLMs.

Results: There was a significant difference between the 3 LLMs in their 5-point Likert scores, with mean values of 4.46 (SD .82), 3.89 (.80), and 3.11 (.72) for ChatGPT-4, Bard, and Bing ($P < .001$). There was a significant difference between the 3 LLMs in the proportion of appropriate answers, with ChatGPT-4 at 85.7%, Bard at 74.3%, and Bing at 25.7% ($P < .001$). The mean FRE scores for ChatGPT-4, Bard, and Bing, were 21.68 (SD 2.78), 42.89 (4.03), and 14.64 (5.09), respectively, with higher scores representing easier readability.

Funding: The research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Correspondence: Jerry Dang, M.D., Ph.D., Digestive Diseases Institute, Cleveland Clinic, Cleveland, OH.

E-mail address: dangj3@ccf.org (J.T. Dang).

https://doi.org/10.1016/j.soard.2024.03.011

Conclusions: LLM-based AI chat models can effectively generate appropriate responses to clinical questions related to bariatric surgery, though the performance of different models can vary greatly. Therefore, caution should be taken when interpreting clinical information provided by LLMs, and clinician oversight is necessary to ensure accuracy. Future investigation is warranted to explore how LLMs might enhance healthcare provision and clinical decision-making in bariatric surgery. (Surg Obes Relat Dis 2024;20:603–608.)

Keywords: Large language models; Chatbot; Artificial intelligence; Bariatric surgery guidelines

The incidence of bariatric and metabolic surgery as interventions for obesity is growing rapidly, with around 507,298 operations recorded by the International Federation for Surgery of Obesity and Metabolic Disorders (IFSO) in 2021 [1]. The diversity of surgical interventions and perioperative considerations introduces a multitude of clinical factors to consider. This underscores the necessity for comprehensive educational materials to inform clinical decision making. Such materials encompass a broad range of knowledge, including intra-operative considerations, management of obesity-related comorbidities, postoperative patient care, and nutritional guidance [2,3].

The vast and dynamic expanse of literature surrounding bariatric surgery offers essential insights, yet its depth and breadth can impose significant demands on time and attention of busy clinicians. With the recent increase in physician burnout due to the strain placed on healthcare systems during the COVID-19 pandemic, the use of online tools for information gathering may be paramount in alleviating the physician workload [4]. Artificial Intelligence-based large language models (LLMs) such as ChatGPT have recently garnered substantial attention, with studies assessing their use in various medical settings, such as writing the United States Medical Licensing Exam, providing recommendations about cardiovascular disease prevention, and responding to patient questions posed on an online forum [5,6]. AI LLMs can assist researchers as well by answering content-specific questions, summarizing texts, and generating bibliographies, outlines, as well as other aids. Moreover, LLMs are able to address a wide range of medical topics with access to knowledge of diseases, treatments, and medical procedures, providing quick access to information or clarification regarding a topic [7]. While AI tools have shown promise in certain fields of medicine, their ability to provide up-to-date information and accurately make clinical decisions in the field of bariatric and metabolic surgery is uncertain.

This study aims to compare the performance of 3 major AI LLMs (ChatGPT, Bard, and Bing) in generating appropriate and clinically accurate responses to common clinical questions relating to bariatric surgery.

Methods

Selection of prompts

The creation of prompts for use in the AI tools involved several pre-existing guidelines pertaining to bariatric and metabolic surgery. Guidelines were included in this study if they met the following criteria: 1) guidelines published after 2010 concerning the perioperative management of patients undergoing or underwent bariatric/metabolic surgery; 2) articles developed by recognized surgical or medical societies; 3) articles developed by committees consisting of recognized experts or affiliates of medical or surgical societies; 4) articles available in full text in the English language. A total of 26 pre-existing bariatric and metabolic guidelines were queried for pertinent clinical questions. We segmented questions into distinct categories as follows: bariatric surgery techniques, complications, weight recurrence, special populations, perioperative management, nutrition, and concurrent surgery. For example, a "bariatric surgery techniques" question included was "Which is more effective for weight loss and treating obesity-related comorbidities: sleeve gastrectomy or gastric bypass?," and the full list of questions is available in Appendix 1. The gathered questions were reviewed to remove duplicates and altered to improve clarity and coherence. The final set of prompts was reviewed and given final approval by all co-authors.

Data generation with AI-based chat models

A total of 36 open-ended clinical questions pertaining to bariatric surgery were queried into 3 AI-based chat models: OpenAI's ChatGPT-4, Microsoft's Bing, and Google's Bard. New chat conversation was created for each question from May 25th, 2023 to May 30th, 2023. Each model generated 36 responses to be reviewed (sample answers available in Appendix 3). An Excel Spreadsheet was used to organize each set of responses, with separate columns used for each AI chat model answer. The responses were rephrased to remove generic introductory phrases, asterisks, and in-text citations. To further assist in the blinding process, the names of the AI chat models were removed and the responses were placed in a random order following each respective question.

A blinded copy of the master spreadsheet was provided to 6 accredited bariatric surgeons in North America (U.S. and Canada) for blinded review. Each surgeon was assigned a grouping of twelve questions to review, and each response was assessed in duplicate.

Outcomes

Our primary outcome was the appropriateness of response provided by each AI LLM, defined as the factual accuracy of the response, comprehensiveness, and consistency of the information provided. Reviewers independently determined an appropriateness rating by assessing each response with to a 5-point Likert scale (Appendix 2). For our analyses, scores of 4 and 5 were deemed appropriate, while scores of 1–3 were considered inappropriate. To assess the readability of the responses, a Flesch Reading Ease (FRE) score was determined for each response by each reviewer along with reasoning behind their appointed scores. The FRE score is calculated based on the formula $206.835 - 1.015 \times$ [words/sentences] $- 84.6 \times$ [syllables/words] and ranges from 0 (Difficult to Read) to 100 (Easy to Read) [8,9]. This score provides an assessment of how easily digestible the content produced by each LLM is. A score of 60–70 is considered "plain English" easily understood by eighth or ninth grade students, whereas lower scores represent more complicated writing, with a score of 30–50 representing "difficult" college-level writing, and 10–30 representing "very difficult" college-graduate level writing.

Statistical analysis

For each LLM's response to a question, Likert scores across the 2 reviewers were combined by rounding up the mean score to the nearest whole number, and then labelled as "appropriate" (4 or 5) or "inappropriate"(1–3). We then determined proportion of appropriate answers per LLM and per question category in percentages. Furthermore, for each LLM, we determined overall mean Likert scores and standard deviation (SD) across all surveyed questions, as well as mean readability scores with SD. Dichotomous variables were compared using the 2x3 chi-square test. Continuous variables were compared with one way analysis of variance (ANOVA) and were provided as mean and standard deviation (SD) or median and interquartile range (IQR). The consistency of Likert scores assigned to each AI LLM response between our evaluators was calculated using a 2-way analysis of variance of intra-class correlation coefficients (ICCs). The level of consistency was based on predetermined ranges for the ICCs: 1) .01 to .20, minor consistency; 2) .21 to .40, fair; 3) .41 to .60, moderate; 4) .75 to .90, good; 5) .90–1.00, excellent. All statistical tests were 2-sided with the threshold for significance set at $P < .05$ and 95% confidence intervals (CI) were provided where applicable. All statistical analyses were performed using *STATA* (StataCorp version 17, College Station, TX).

Results

The number and proportion of appropriate answers provided by each AI model are presented in Table 1. ChatGPT-4 demonstrated the highest performance with 30 (85.7%) of its answers deemed appropriate, while Bing and Bard provided 9 (25.7%) and 26 (74.3%) appropriate answers, respectively ($P < .001$). A significant difference was also noted in the mean Likert scores between the AI models, with ChatGPT-4 receiving the highest mean score of 4.46 (SD .82) ($P < .001$). Likert scores for each response can be found in Appendix 1. Analyzing the average word count of the responses generated by each AI model, Bard-generated responses with highest median word count of 269 (IQR: 248–332), with ChatGPT-4 and Bing providing responses with median words counts of 257 (IQR: 226–299) and 100 (IQR: 83–115) ($P < .001$) respectively. The average ICC for the included studies was .904. The evaluators displayed a very high level of consistency given that the lowest ICC was .845.

A subgroup analysis was also conducted based on question category, which can be seen in Table 2. ChatGPT-4 continued to display the highest mean Likert score in questions related to surgery techniques, perioperative management, as well as concurrent surgery. Bard presented the highest mean Likert scores for weight recurrence questions while Bing displayed the lowest average scores across all subgroups.

Tables 3 and 4 outlines the FRE scores and Flesch-Kincaid Grade (FKG) scores for each AI model and their responses according to category. Bard received the highest mean FRE score across all categories with 42.89 points (SD: 4.03) compared to Bing with 14.64 points (SD: 5.09) and ChatGPT-4 with 21.68 points (SD: 2.78) ($P < .001$).

Discussion

This is the first study to evaluate and compare the quality and readability of 3 AI-chat-generated responses to clinical questions related to bariatric and metabolic surgery. Our study found a significant difference between the 3 LLMs in their ability to appropriately answer clinical questions

Table 1
Number of appropriate answers from AI models

AI model	ChatGPT-4	Bing	Bard	*P* value
5-point Likert Scale, Mean (SD)	4.46 (.82)	3.11 (.72)	3.89 (.80)	<.001
Appropriate Answers, N (%)	30 (85.7%)	9 (25.7%)	26 (74.3%)	<.001
Word Count, Median (IQR)	257 (226–299)	100 (83–115)	269 (248–332)	<.001

IQR = interquartile range; AI = artificial intelligence.

Table 2
Average Likert score based on question categories

AI model	ChatGPT-4, mean (SD)	Bing, mean (SD)	Bard, mean (SD)	*P* value
Surgery techniques	4.66 (.52)	2.67 (.82)	3.5 (1.04)	.003
Complications	4.2 (1.10)	2.8 (.45)	3.2 (.84)	.056
Weight Recurrence	3.67 (.58)	3 (0)	4 (0)	.027
Special Population	4.22 (1.09)	3.22 (.83)	4 (.71)	.063
Perioperative management	4.86 (.38)	3.14 (.69)	4.29 (.76)	<.001
Surgery and nutrition	4.66 (.58)	4 (0)	4.33 (.58)	.30
Concurrent surgery	5 (0)	3.5 (.71)	4 (0)	.074

SD = standard deviation.

related to bariatric surgery, with ChatGPT-4 scoring the largest proportion of appropriate answers. Bing, in particular, had very poor performance, providing mostly inappropriate responses. The accuracy of answers to questions regarding surgical technique, weight recurrence, and perioperative management also varied significantly, with ChatGPT-4 consistently providing the most accurate information. In terms of readability, Bard's answers were easiest to read on average, although it was still considered "Difficult" with scores ranging from 30 to 50 on the FRE scale [8,9]. Both ChatGPT-4 and Bing offered answers that were "Very Difficult" to read and were at a college-graduate reading level based on the FRE. Bard's answers were the most succinct whereas ChatGPT-4 and Bing provided longer, yet comparably extensive, answers.

The growing demand for bariatric surgery, fueled by the rising prevalence of obesity, has underscored its safety and efficacy in providing a long-term solution for weight reduction [10–12]. This increased popularity has amplified the demand for easily accessible resources to aid clinical decision-making. However, individual research manuscripts or lengthy clinical practice guidelines can be time-consuming and challenging to grasp for general practioners. One potential application for AI-based chat models is as a reliable and effective resource for physicians seeking prompt information sourced from clinical guidelines. By serving as a repository of consolidated clinical recommendations, AI LLMs could streamline the dissemination of latest and the most accurate evidence in bariatric surgery, potentially increasing adherence to novel recommendations. Another application of LLMs lies in their capacity to address patient questions undergoing bariatric surgery. Patients often turn to online resources including social media as a resource for information related to bariatric surgery, however, the quality and reliability of these sources have been questioned [13–17]. Given the variable quality of information from online and social media resources, LLMs stand out for their ability to provide logical and tailored responses to questions related to bariatric surgery. In addition, recent studies have demonstrated that ChatGPT-4 responds to patient medical questions with empathetic and accurate answers [18,19]. Over the last few years, the surge in electronic messaging has led to increased after-hours work for physicians [20–22]. As such, incorporation of a LLMs in this context could mitigate this workload and reduce rates of physician burnout. The incorporation of an AI-based chat model in a bariatric clinic, for instance, may help address patient questions in a timely manner, and may be especially advantageous for patients who rely on messaging and telemedicine due to demographic or socioeconomic barriers. Overall, our study demonstrates that AI-based chat models exhibit promise in potentially enhancing physician workflow by providing meaningful answers to clinical queries in the field of bariatrics.

While these findings are promising, the study's limitations must be acknowledged. First, subjectivity might be by expert raters and the chosen evaluation criteria. Second, with respect to AI LLMs being used as conversational agents for patients, important considerations include chat monitoring,

Table 3
Average Flesch Reading Ease Score based on question categories

AI model	ChatGPT-4, mean (SD)	Bing, mean (SD)	Bard, mean (SD)	*P* value
Surgery techniques	26.05 (6.10)	23.38 (4.51)	41.93 (14.98)	.015
Complications	27.92 (5.75)	17.16 (11.97)	41.86 (7.21)	<.001
Weight Recurrence	32.2 (1.14)	14.77 (13.91)	46.47 (1.25)	.022
Special Population	12.14 (4.00)	14.49 (14.81)	36.99 (7.45)	<.001
Perioperative management	18.57 (9.61)	12.54 (6.07)	43.93 (10.11)	<.001
Surgery and nutrition	25.97 (1.16)	18.37 (2.59)	43.97 (7.31)	.003
Concurrent surgery	8.9 (3.5)	1.75 (1.75)	45.1 (4.1)	.005
All questions	21.68 (2.78)	14.64 (5.09)	42.89 (4.03)	<.001

SD = standard deviation.

Table 4
Average Flesch–Kincaid Grade Level based on question categories

AI model	ChatGPT-4, mean (SD)	Bing, mean (SD)	Bard, mean (SD)	*P* value
Surgery techniques	14 (.74)	14 (1.44)	11 (2.23)	.004
Complications	13 (.69)	15 (2.20)	11 (.96)	<.001
Weight Recurrence	12 (.45)	16 (2.12)	10 (.94)	.016
Special Population	16 (1.83)	16 (2.20)	12 (1.73)	<.001
Perioperative management	14 (1.23)	15 (1.33)	10 (1.69)	<.001
Surgery and nutrition	13 (.52)	14 (.41)	10 (1.84)	.043
Concurrent surgery	17 (.7)	17 (.45)	10 (.55)	.005
All questions	14 (.45)	15 (.73)	11 (.57)	<.001

establishing escalation pathways to clinicians, trust and transparency, privacy and cybersecurity, and protection from commercially motivated data sharing or marketing [23]. Furthermore, all AI models in our study provided some answers that our reviewers considered inappropriate; for instance, ChatGPT-4 provided inappropriate answers to 14.3% of questions, and Bing answered 74.3% of questions inappropriately. This relates to inherent issues of LLMs, which is the possibility to "hallucinate", that is, to produce inaccurate or nonsensical information [24,25]. This is notable in our study, as Bing specifically had a high proportion of inappropriate answers, despite being based on the same GPT-4 LLM as ChatGPT. The stark difference in performance between ChatGPT and Bing is potentially due to the Bing AI incorporating search engine-derived data from the Bing search engine. This allows Bing to access up-to-date information and evolve its answers; however, it also means that it is prone to accessing incorrect and irrelevant information, introducing inaccuracies into its responses.

While physicians are not exempt from making errors, with even the most skilled and well-meaning professionals occasionally provide incorrect advice, it is crucial to maintain ongoing training and refinement of AI models to improve the precision of their responses [26]. The refinement process should harness the burgeoning volume of robust data and integrate extensive fine-tuning supervised by a diverse team of experts to maximize accuracy. A clinician's oversight remains essential to validate the accuracy of AI-generated advice in bariatric surgery until it reaches a level of reliability comparable to that of a clinical expert. Finally, our study found that the AI LLMs provided answers that were overall considered "difficult" to read. Bard was, on average, easiest to read, and this may be attributed to its shorter answers given that word count and the number of sentences integral factors in the FRE formulas. As patient education levels are important factors in health literacy, counselling patients on how to properly use the AI model to fit their needs and providing clear direction on when to touch base with a clinician if answers remain unclear, is imperative. Conversely, the present reading complexity of all 3 AI models would be appropriate for general practioners. It is noteworthy that AI LLMs have the capacity to rephase responses into more straightforward or layman's terms upon request; this feature was not examined in our current study but poses an intriguing avenue for furture research to explore.

Conclusion

Our findings suggest that AI-based chat models, particularly those based on the GPT-4 architecture like ChatGPT-4, hold significant potential to serve as an effective tool in delivering timely and relevant information to questions arising in bariatric surgery. While AI models are not intended to replace human expertise, they could be integrated to augment clinical practice, enhancing efficiency and accessibility of information for both healthcare professionals and patients. Continuous advancements in AI technology and further empirical studies will shape the role of these tools in bariatric surgery research and applications, such as in the development of domain-specific models tailored for bariatric surgery. Future research should further investigate and build upon the capacity of AI to support clinical decision making in bariatric surgical practice.

Disclosure

All authors disclose no relevant conflicts of interest.

Supplementary data

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.soard.2024.03.011.

References

[1] Picot J, Jones J, Colquitt JL, et al. The clinical effectiveness and cost-effectiveness of bariatric (weight loss) surgery for obesity: a systematic review and economic evaluation. Health Technol Assess 2009;13:1–190. 215-357, iii-iv.

[2] Pratt JSA, Browne A, Browne NT, et al. ASMBS pediatric metabolic and bariatric surgery guidelines, 2018. Surg Obes Relat Dis 2018;14:882–901.

[3] Heber D, Greenway FL, Kaplan LM, et al. Endocrine and nutritional management of the post-bariatric surgery patient: an Endocrine Society Clinical Practice Guideline. J Clin Endocrinol Metab 2010;95:4823–43.

[4] Shanafelt TD, West CP, Dyrbye LN, et al. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the COVID-19 pandemic. Mayo Clin Proc 2022;97:2248–58.

[5] Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023;329:842–4.
[6] Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589–96.
[7] Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Available from: http://commoncrawl.org/. Accessed July 19, 2023.
[8] Flesch R. A new readability yardstick. J Appl Psychol 1948;32:221–33.
[9] Kincaid J, Fishburne R, Rogers R, et al. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training. Available from: https://stars.library.ucf.edu/istlibrary/56. Accessed August 9, 2023.
[10] Xia Q, Campbell JA, Ahmad H, et al. Bariatric surgery is a cost-saving treatment for obesity-a comprehensive meta-analysis and updated systematic review of health economic evaluations of bariatric surgery. Obes Rev 2020;21:e12932.
[11] Alalwan AA, Friedman J, Park H, et al. US national trends in bariatric surgery: a decade of study. Surgery 2021;170:13–7.
[12] Lo Menzo E, Szomstein S, Rosenthal RJ. Changing trends in bariatric surgery. Scand J Surg 2015;104:18–23.
[13] Scarano Pereira JP, Martinino A, Manicone F, et al. Bariatric surgery on social media: a cross-sectional study. Obes Res Clin Pract 2022;16:158–62.
[14] Athanasiadis DI, Roper A, Hilgendorf W, et al. Facebook groups provide effective social support to patients after bariatric surgery. Surg Endosc 2021;35:4595–601.
[15] Batar N, Kermen S, Sevdin S, et al. Assessment of the quality and reliability of information on nutrition after bariatric surgery on YouTube. Obes Surg 2020;30:4905–10.
[16] Corcelles R, Daigle CR, Talamas HR, et al. Assessment of the quality of internet information on sleeve gastrectomy. Surg Obes Relat Dis 2015;11:539–44.
[17] Koball AM, Jester DJ, Pruitt MA, et al. Content and accuracy of nutrition-related posts in bariatric surgery Facebook support groups. Surg Obes Relat Dis 2018;14:1897–902.
[18] Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33:1790–6.
[19] Ayers JW, Zhu Z, Poliak A, et al. Evaluating artificial intelligence responses to public health questions. JAMA Netw Open 2023;6:e2317517.
[20] Zulman DM, Verghese A. Virtual care, telemedicine visits, and real connection in the era of COVID-19: unforeseen opportunity in the face of adversity. JAMA 2021;325:437–8.
[21] Holmgren AJ, Downing NL, Tang M, et al. Assessing the impact of the COVID-19 pandemic on clinician ambulatory electronic health record use. J Am Med Inform Assoc 2022;29:453–60.
[22] Tai-Seale M, Dillon EC, Yang Y, et al. Physicians' well-being linked to in-basket messages generated by algorithms in electronic health records. Health Aff (Millwood) 2019;38:1073–8.
[23] McGreevey JD, Hanson CW, Koppel R. Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. JAMA 2020;324:552–3.
[24] Rohrbach A, Hendricks LA, Burns K, et al. Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018; 2018. p. 4035–45. Association for Computational Linguistics; Brussels, Belgium.
[25] Xiao Y, Wang WY. On hallucination and predictive uncertainty in conditional language generation. In: EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference; 2021. p. 2734–44 Association for Computational Linguistics (virtual conference).
[26] Fine-tuning - OpenAI API. Available from: https://platform.openai.com/docs/guides/fine-tuning. Accessed July 29, 2023.