The Oncologist, 2024, **29**, 407–414
https://doi.org/10.1093/oncolo/oyae009
Advance access publication 3 February 2024
Original Article

Comparison of Large Language Models in Answering Immuno-Oncology Questions: A Cross-Sectional Study

Giovanni Maria Iannantuono[1,‡], Dara Bracken-Clarke[2,‡], Fatima Karzai[1], Hyoyoung Choo-Wosoba[3], James L. Gulley[2], Charalampos S. Floudas[*,2,]

[1]Genitourinary Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[2]Center for Immuno-Oncology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[3]Biostatistics and Data Management Section, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

[*]Corresponding author: Charalampos S. Floudas, MD, Center for Immuno-Oncology, Center for Cancer Research, National Cancer Institute, Building 10, Room 7N240A, 10 Center Drive, Bethesda, MD 20892, USA. Tel: +1 240 858 3032. Email: charalampos.floudas@nih.gov
[‡]Contributed equally.

Abstract

Background: The capability of large language models (LLMs) to understand and generate human-readable text has prompted the investigation of their potential as educational and management tools for patients with cancer and healthcare providers.

Materials and Methods: We conducted a cross-sectional study aimed at evaluating the ability of ChatGPT-4, ChatGPT-3.5, and Google Bard to answer questions related to 4 domains of immuno-oncology (Mechanisms, Indications, Toxicities, and Prognosis). We generated 60 open-ended questions (15 for each section). Questions were manually submitted to LLMs, and responses were collected on June 30, 2023. Two reviewers evaluated the answers independently.

Results: ChatGPT-4 and ChatGPT-3.5 answered all questions, whereas Google Bard answered only 53.3% ($P < .0001$). The number of questions with reproducible answers was higher for ChatGPT-4 (95%) and ChatGPT3.5 (88.3%) than for Google Bard (50%) ($P < .0001$). In terms of accuracy, the number of answers deemed fully correct were 75.4%, 58.5%, and 43.8% for ChatGPT-4, ChatGPT-3.5, and Google Bard, respectively ($P = .03$). Furthermore, the number of responses deemed highly relevant was 71.9%, 77.4%, and 43.8% for ChatGPT-4, ChatGPT-3.5, and Google Bard, respectively ($P = .04$). Regarding readability, the number of highly readable was higher for ChatGPT-4 and ChatGPT-3.5 (98.1%) and (100%) compared to Google Bard (87.5%) ($P = .02$).

Conclusion: ChatGPT-4 and ChatGPT-3.5 are potentially powerful tools in immuno-oncology, whereas Google Bard demonstrated relatively poorer performance. However, the risk of inaccuracy or incompleteness in the responses was evident in all 3 LLMs, highlighting the importance of expert-driven verification of the outputs returned by these technologies.

Key words: large language models; artificial intelligence; immuno-oncology; ChatGPT; Google Bard.

Implications for Practice

Several studies have recently evaluated whether large language models may be feasible tools for providing educational and management information for cancer patients and healthcare providers. In this cross-sectional study, we assessed the ability of ChatGPT-4, ChatGPT-3.5, and Google Bard to answer questions related to immuno-oncology. ChatGPT-4 and ChatGPT-3.5 returned a higher proportion of responses, which were more accurate and comprehensive, than those returned by Google Bard, yielding highly reproducible and readable outputs. These data support ChatGPT-4 and ChatGPT-3.5 as powerful tools in providing information on immuno-oncology; however, accuracy remains a concern, with expert assessment of the output still indicated.

Introduction

Large language models (LLMs) are a recent breakthrough in the domain of generative artificial intelligence (AI).[1] Generative AI includes technologies based on "natural language processing" (NLP) which uses computational linguistics and deep learning (DL) algorithms to enable computers to interpret and generate human-like text.[2] Large language models are complex systems trained on large quantities of text data which are able to create new content in response to prompts such as text, images, or other media.[3] This versatility has led to the investigation of their potential applications in the field of medicine and healthcare in light of their self-evident potential benefits in these domains.[4] Indeed, the availability of user-friendly tools able to provide detailed, accurate, and current information would be crucial in promoting patient and healthcare providers' education and awareness, particularly in the case of complex health conditions like cancer.[5]

Received: 10 November 2023; Accepted: 11 January 2024.

Thus far, many studies have assessed the potential of ChatGPT, an advanced LLM based on a generative pre-trained transformer (GPT) architecture, for providing screening and/or management information in solid tumors.[6] Following the rollout of ChatGPT, more LLMs trained on different data were released, expanding the selection of these new AI-based tools. Consequently, an increasing number of studies are investigating and comparing the potential ability of ChatGPT with other LLMs as easy-to-use interfaces to gather information related to a specific cancer-related topic.[7] Particularly, LLMs were assessed for their capability of both providing accurate and relevant responses to specific questions and writing in a comprehensible and coherent natural language.[7] So far, initial evidence suggests a possible role of these technologies as "virtual assistants" for healthcare professionals and patients in providing information about cancer, which is unfortunately counterbalanced by a significant error rate.[7] Notably, while representing a remarkable achievement in computer science, the need for precision, accuracy, and legal responsibility in the field of medicine and healthcare represents a significant obstacle to their implementation, especially outside of trials and close expert monitoring.[7]

The past several years have seen profound changes in the field of immuno-oncology (IO). The advent of immune-checkpoint inhibitors (ICIs) has paved the way toward a new era in cancer treatment, enhancing the chance of long-term survival in patients with metastatic disease, and providing new treatment options in earlier-stage settings.[8] Presently, an increasing number of patients with cancer are either candidates or already receiving ICIs or other immunotherapies, subject to both the enormous potential benefits but also the immune-related adverse events that may be caused by these treatments.[9] In this context, LLMs may represent a valid tool for healthcare professionals and patients (and their caregivers) receiving these treatments. Therefore, we sought to assess and compare the ability of 3 prominent LLMs to provide educational and management information in the IO field.

Materials and Methods

Large Language Models

In this cross-sectional study, we compared the performance of 3 LLMs: ChatGPT-3.5,[10] ChatGPT-4,[10] and Google Bard.[11] ChatGPT is an LLM based on the GPT architecture and developed by OpenAI, a company based in San Francisco (USA). ChatGPT is built upon either GPT-3.5 or GPT-4; the former is freely available to all the users, whereas the latter is an advanced version with additional features and provided under the name "ChatGPT Plus" to paid subscribers.[10] Google Bard is based on the Pathways Language Model (PaLM) family of LLMs, developed by GoogleAI.[11]

Questions and Responses' Generation

We generated 60 open-ended questions based on our clinical experience covering 4 different domains of IO including "mechanisms" (of action), "indications" (for use), "toxicities," and "prognosis" (Supplementary Material A). Questions were manually and directly submitted to the web chat interfaces of the 3 abovementioned LLMs on June 30, 2023 and responses were collected (Supplementary Material B). We assessed the reproducibility, accuracy, relevance, and readability (Table 1) of responses provided by each LLM. Two reviewers (G.M.I. and D.B.C.) rated the answers independently. Before submitting the questions to the LLMs, reviewers created a sample response for each question to take as a reference for assessing accuracy and relevancy during the rating process. Furthermore, reviewers were blinded to the LLM being assessed. Inconsistencies between the reviewers were discussed with an additional reviewer (C.S.F.) and resolved by consensus. Cohen's kappa coefficient was calculated to evaluate inter-rater reliability during the rating process.[12]

First, we assessed the ability of each LLM to provide reproducible responses. Therefore, each individual question was submitted 3 times on each LLM. In the case of non-reproducible answers, questions were not considered for further analysis. Subsequently, the accuracy, relevance, and

Table 1. Definitions of the outcomes.

Outcomes	Definitions	Score
Answer returned	The ability of LLM to return a meaningful answer to each instance of the question submitted, rather than returning an error or declining to return an answer, independent of the accuracy of this response	Recorded as Boolean True/False
Reproducibility	The ability of LLM to return a generally similar series of answers across the 3 separate queries with no fundamental differences or inconsistencies between these 3 answers	Recorded as Boolean True/False
Accuracy	The ability of LLM to provide accurate and correct information addressing the question asked and returning all major or critical points required in such an answer. Response not adversely marked for extraneous or irrelevant information here—as long as this information was correct	Recorded numerically from 1 to 3
Readability	The ability of LLM to return comprehensible and coherent natural language text in English, including appropriate syntax, formatting, and punctuation, independent of the accuracy of this response	Recorded numerically from 1 to 3
Relevance	The ability of LLM to return information that was relevant and specific to the question asked or immediately adjacent topics without extraneous, unrequested, or tangential information. Accuracy was not specifically assessed here, although the result was adversely marked if the response included immaterial information while neglecting to address the specific question asked	Recorded numerically from 1 to 3

For scoring of Relevance, the answer returned was not adversely marked for any included disclaimers to the effect that the LLM cannot provide medical advice and any such advice should be sought from a clinician or that anyone with a cancer diagnosis and/or receiving systemic therapy with potential toxicity should contact their treating clinician/s. This was deemed to represent appropriate and medically sound advice and not to be irrelevant or extraneous material.

readability of responses deemed reproducible were assessed using a 3-point scale (Table 2; Fig. 1). Reviewers graded the accuracy of answers according to available information as of 2021, as the training datasets of ChatGPT are updated on September 2021. Finally, word- and character-counts were calculated for each answer.

Statistical Analyses

Categorical variables were presented with proportions and numeric variables as measures of central tendency. Comparisons between categorical variables were performed with 2-sided generalized Fisher's exact tests for testing any potential differences in these 3 LLMs. In the case of numeric continuous variables, a Kruskal-Wallis test was used. Statistical tests were not performed within each of the 4 domains, but rather were performed only to evaluate overall performance by combining those 4 domains, due to insufficient sample sizes within each domain (ie, only up to 15 available observations). All statistical results should be interpreted as exploratory; all statistical analyses were performed and all plots generated using R version 4.2.2 (The R Foundation for Statistical Computing, 2022). This study was conducted in accordance with Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.[13]

Results

Assessment of inter-rater reliability with Cohen's kappa during the rating process demonstrated "strong" to "near perfect" agreement between reviewers (Table 3). ChatGPT-3.5 and ChatGPT-4 provided at least one response to all questions (60 [100%]), while Google Bard responded only to 32 (53.3%) queries ($P < .0001$). Specifically, the percentages of responses provided by Google Bard were different across the 4 domains, with better performances in the "mechanisms" (14 [93.3%]) and "prognosis" domains (13 [86.7%])

Table 2. Definitions of the scoring system.

	Score		
	1	2	3
Accuracy	Fundamentally inaccurate or incorrect information, including critical errors, omissions and/or entirely incorrect treatment advice	Partially correct and accurate information, including non-critical errors and/or omitting relevant information or failing to provide specific guideline advice	Fully accurate and correct information, answering the specific question asked with no significant errors or omissions
Relevance[a]	Irrelevant and/or entirely tangential material, not addressing the specific question asked	Generally relevant material although including significant extraneous and/or tangential information	Relevant and focused information directly addressing the question asked, including an appropriate expansion on the relevant topic
Readability	Incoherent, unintelligible and/or garbled text, ± severely misformatted and/or oxymoronic material resulting in compromised legibility	Generally coherent and intelligible material with significant formatting and/or parsing errors	Fully coherent, well-parsed and constructed material, easily and clearly intelligible

[a] Inclusion of a disclaimer that the answer was provided by an AI/LLM and cannot be taken as medical advice and/or that any information or questions should also be addressed to a qualified medical practitioner was not scored negatively—as this represents a legitimate and appropriate legal disclaimer.

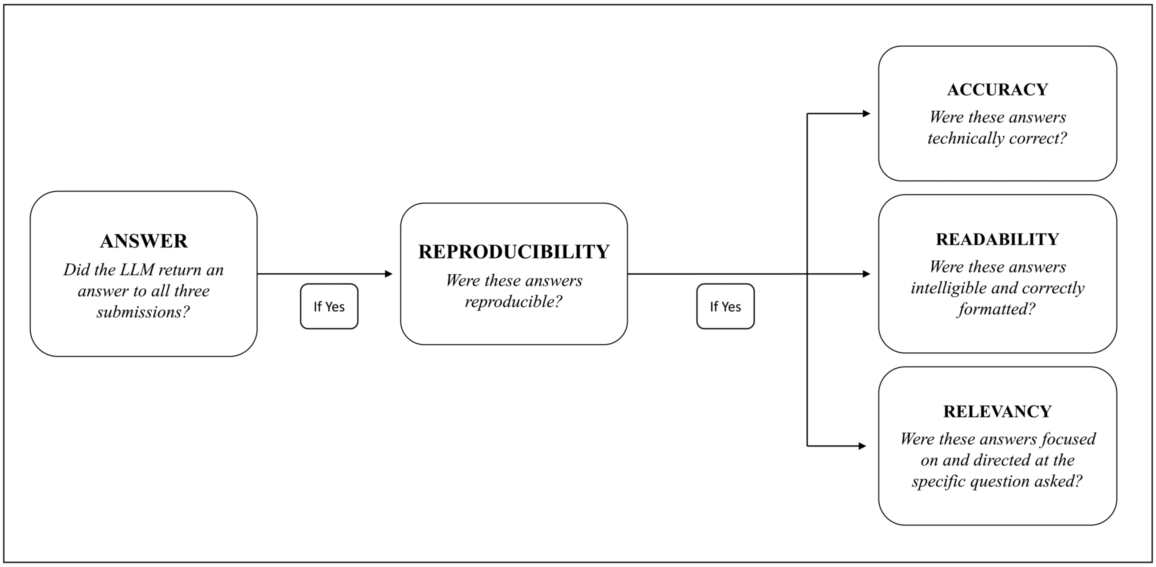
ANSWER
Did the LLM return an answer to all three submissions?
If Yes
REPRODUCIBILITY
Were these answers reproducible?
If Yes
ACCURACY
Were these answers technically correct?
READABILITY
Were these answers intelligible and correctly formatted?
RELEVANCY
Were these answers focused on and directed at the specific question asked?

Figure 1. Flowchart of the rating process for each triplet of responses.

compared to the "indications" (5 [33.3%]), and "toxicities" (0 [0%]) domains. Regarding reproducibility, the numbers of questions with reproducible answers were similar between ChatGPT-3.5 and ChatGPT-4 (53 [88.3%] and 57 [95%], respectively), while it was lower (16 [50%]) for Google Bard ($P < .0001$). Although ChatGPT-3.5 and ChatGPT-4 performed similarly across all domains, ChatGPT-4 achieved 100% reproducible responses in 2 domains ("mechanisms" and "indications") in which ChatGPT-3.5 achieved only 86.7%. Google Bard was variably capable and accurate across the different sections. Despite a significant number of answers deemed reproducible in the "mechanisms" (6 [40%]) and "prognosis" (9 [60%]) sections, a poor performance was observed in the "indications" (1 [6.7%]) and "toxicities" (0 [0%]) domains (Fig. 2). In terms of accuracy, the numbers of answers deemed fully correct were 31 (58.5%), 43 (75.4%), and 7 (43.8%) for ChatGPT-3.5, ChatGPT-4, and Google Bard, respectively ($P = .03$). Furthermore, regarding relevancy, the numbers of responses deemed highly relevant were 41 (77.4%), 41 (71.9%), and 7 (43.8%) for ChatGPT-3.5, ChatGPT-4, and Google Bard, respectively ($P = .04$). Readability was deemed optimal across all 3 LLMs. However, the numbers of highly readable answers were greater for ChatGPT-3.5 and ChatGPT-4 (52 [98.1%] and 57 [100%]) compared to Google Bard (14 [87.5%]) ($P = .02$; Fig. 3). The median numbers of words and their corresponding ranges for the responses provided by ChatGPT-3.5, ChatGPT-4, and Google Bard were 297 (197-404), 276 (139-395), and 290.5 (12-424), respectively ($P = .06$). Finally, the median numbers of characters and their corresponding ranges were 1829 (1119-2470), 1589 (854-2233), and 1532 (75-2070), respectively ($P < .0001$).

Table 3. Cohen's kappa coefficient for inter-rater reliability between the reviewers during the selection process.

Domain	Cohen's kappa	Agreement (%)	Responses (no. of)
Reproducibility	0.912	97.4	152
Accuracy	0.889	94.4	126
Readability	1	100	126
Relevancy	0.868	94.4	126

Discussion

In recent decades, significant effort has been made to harness the potential of AI in medicine and healthcare.[14] Artificial intelligence can be defined as "the science and engineering of making intelligent machines, especially intelligent computer programs."[15] It is composed of multiple subfields, based on different algorithms and principles, including knowledge representation, machine learning (ML), DL, and NLP.[2,16] Specifically, NLP uses computational language and DL to enable computers to understand text in the same way as humans.[2] Recent progress in NLP has led to major breakthroughs in the field of generative AI, as evidenced by the advent of LLMs.[3] These can recognize, summarize, and generate novel content using statistical connections between letters and words. Indeed, LLMs can also be considered as "few shot

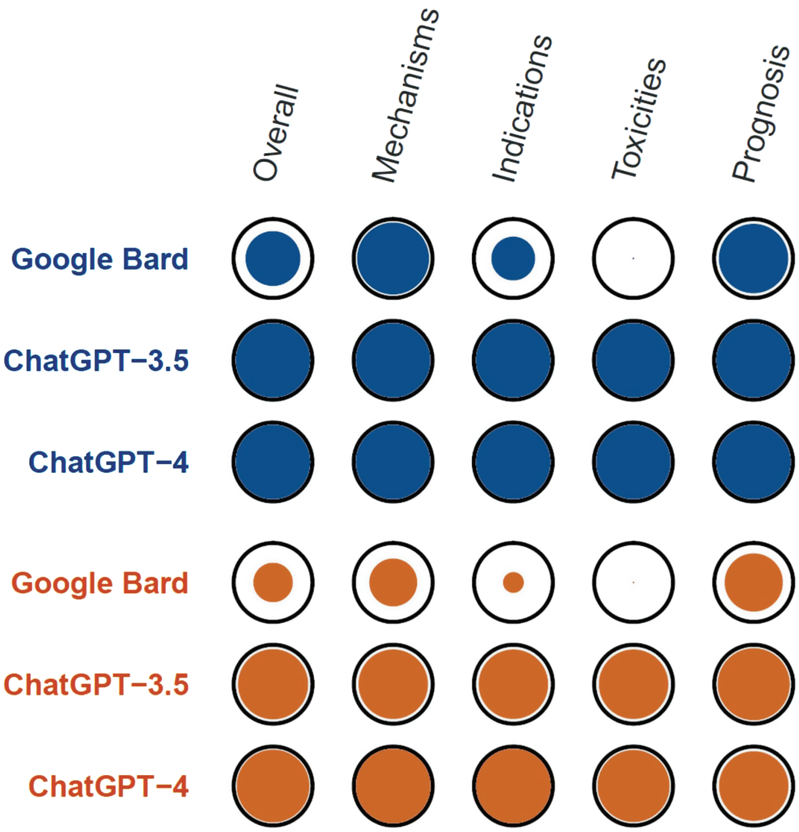

Figure 2. Spot matrix of the percentages of the answered questions (Blue) and reproducible responses (Orange) for each LLM. Color volume is directly proportional to percentage with the outer black circle representing 100%. Corresponding numeric data are available in Supplementary Material C.

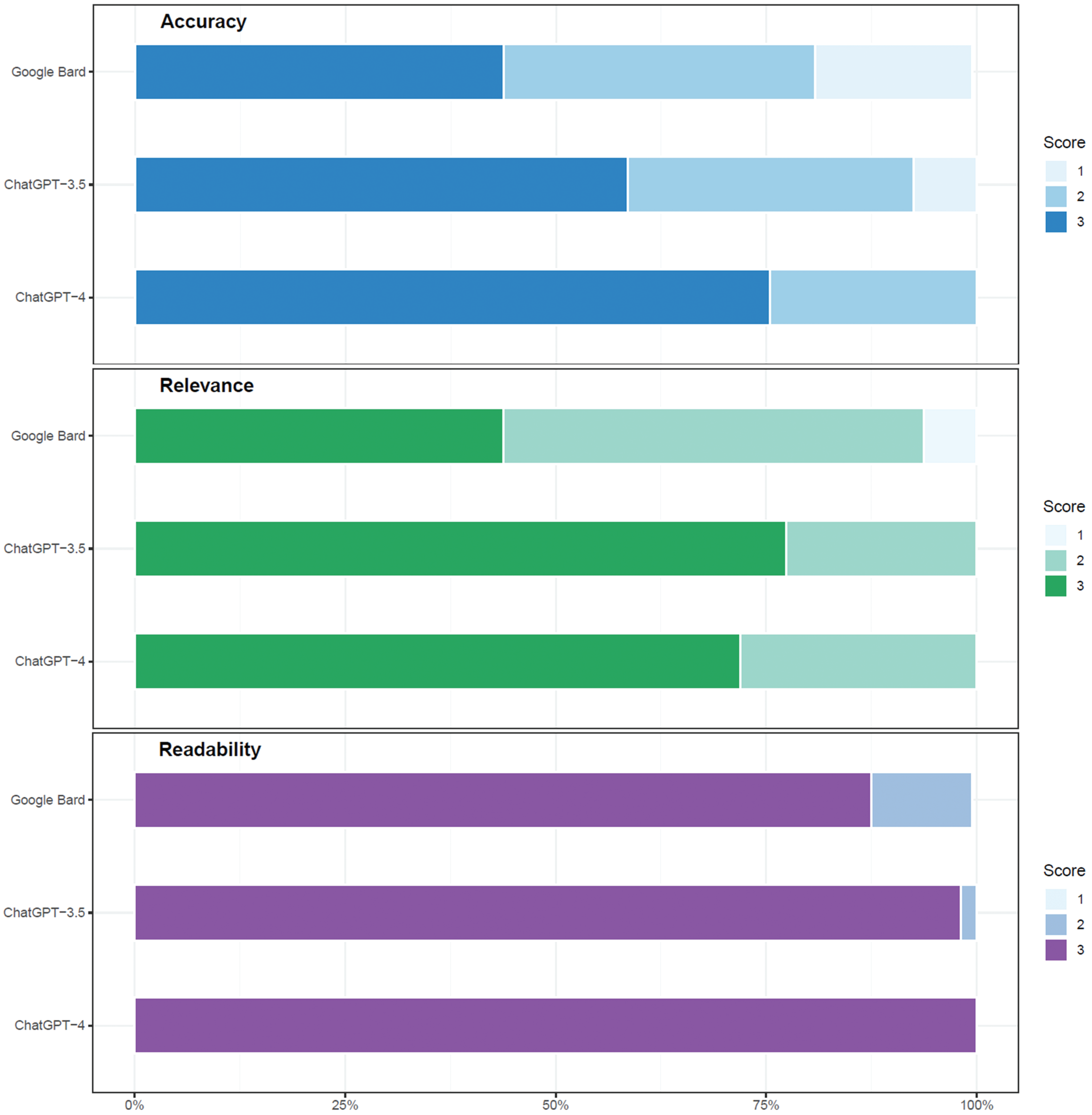

Figure 3. Bar plot of the results (accuracy, readability, and relevance) for all 3 LLMs. This plot was based only on the questions evaluable for accuracy, readability, and relevance. Corresponding numeric data are available in Supplementary Material C.

learners" due to their ability to readily adapt to new domains with few information after being trained.[17]

Over the last year, the release of ChatGPT[10] has attracted considerable attention, which only increased following the release of other LLMs such as Google Bard,[11] Bing AI[18] and, Perplexity.[19] The remarkable adaptability of these AI-based technologies to a broad and extensive range of disciplines was immediately apparent following their introduction.[20] This is also evidenced by the rapid publication of large numbers of studies designed to investigate their role in multiple and diffuse fields, including medicine and healthcare. Initial data have demonstrated LLMs to be highly applicable to the field of cancer care, especially in providing information about the screening and/or management of specific solid tumors.[7] However, to the authors' knowledge, their potential role in the field of IO has not yet been investigated, despite the rapidly expanding knowledge in all the aspects of IO (basic, translational, and clinical research) and the large number of patients with cancer currently receiving immunotherapy.[8,9] Of note, the protean nature of IO toxicities (in both chronology and time course), especially when given in combination with another agent (eg, cytotoxic chemotherapy or tyrosine-kinase inhibitor therapy) whereby almost any symptom *may* relate, at least in part, to the IO agent is likely to complicate the use of LLMs in this field.[21]

Therefore, we performed a cross-sectional study aimed for the first time at assessing the potential of 3 prominent LLMs in answering questions about the field of IO. Our results demonstrated that ChatGPT-4 and ChatGPT-3.5 were able to answer most of the IO-related questions with excellent accuracy and relevance. In contrast, the performance of Google Bard was comparatively poorer, as shown by a lower number of both answered questions and the reproducibility/accuracy of these responses, compared to the other 2 LLMs. All 3 LLMs were able to provide highly readable responses, highlighting the power of these generative AI technologies in providing human-readable text. ChatGPT (both v3.5 and v4) clearly demonstrated their potential as a "virtual assistant" for both clinicians and patients or caregivers. ChatGPT (both v3.5 and, especially, v4) has also demonstrated remarkable acumen in both diagnosing and providing management plans for IO toxicities. It has also proved highly effective in suggesting evidence-based and licensed indications for IO therapy, either alone or in combination. Additionally, it has demonstrable efficacy in providing background information on IO drug mechanisms and disease prognoses in generally comprehensible text without excess jargon, although often with a lack of sources and broken or inaccurate references.

However, the results of this study also highlight the differing performance of various LLMs across topics and specific tasks (Table 4), as this demonstrates significant variability. In our study, ChatGPT is demonstrated to be a powerful tool when applied to the field of IO, particularly in comparison to Google Bard. Similar results were also reported in another recently published study assessing these 3 LLMs in a different cancer-related topic. Specifically, Rahsepar et al reported the results of a study investigating the ability of ChatGPT-3.5, ChatGPT-4, and Google Bard in answering questions related to lung cancer screening and prevention.[32] As in our study, ChatGPT achieved a superior performance to Google Bard. However, the available evidence suggests that the LLM developed by OpenAI is not always accurate, as shown by the results of other studies investigating medical/healthcare topics other than cancer (Table 4). In the studies published by Seth et al, Zúñiga Salazar et al, and Dhanvijay et al, Google Bard performed better in comparison to ChatGPT in non-cancer domains, likely clarifying a potential role for this LLM.[23,24,31] Furthermore, the results of the study by Al-Ashwal et al showed a better performance for Bing AI in answering questions related to drug-drug interactions in comparison to the other LLMs.[22] Therefore, it is essential to compare the performance of different LLMs since their abilities may vary based on both task and domain.

In addition, despite the promising results of our study and its unequivocal efficacy in synthesizing and evaluating textual data, the potential of ChatGPT for error and hallucination remains.[33] The occurrence of "hallucinations" is one of the greatest obstacles to the routine clinical application of LLMs. While potentially tolerable in other domains, this is a critical issue in medicine and the biomedical sciences due to its potential to directly impact patient care. In addition, it must be noted that the datasets on which these models were trained were (i) confidential and proprietary (thus impossible to assess for data quality or bias), (ii) not specifically selected ab initio for addressing biomedical issues, and (iii) only valid up to September 2021 (thus lacking up to date information—a major issue in so rapidly evolving a field as medicine in general and IO in particular).[10,33] Therefore, expert assessment of LLMs' output remains a prerequisite for their clinical use.[34] Following on from these points, the use of LLMs in medicine and healthcare raises profound and difficult questions regarding medical malpractice liability (identify who is responsible if LLMs' recommendations cause harm to patients), intellectual property (if LLMs produce materials similar to scientific or academic studies, this could lead to intellectual property rights issues), and patient data privacy (ensure that patients' data are fully anonymized and protected from possible violations).[35] As a result, it will be essential to develop a regulatory framework in the next future to ensure healthcare professionals and patients use LLMs without risks.

Open-source LLMs trained on specific biomedical datasets in order to accomplish pre-specified tasks offer a potential solution and alternative paradigm. BioGPT, a cutting-edge LLM with a user-friendly interface developed for the biomedical field, represents an excellent example of this.[36] BioGPT shares the same architecture as OpenAI's GPT models but was trained on information derived from the biomedical literature. It has demonstrated excellent performance in several tasks, including text generation and categorization, due to its extensive pre-training on massive biomedical datasets.[36] Further studies to investigate the utility and performance of LLMs developed on biomedical data, with comparison to those LLMs presently available, are, thus, required.

Limitations

Our study has some limitations that need to be mentioned. Firstly, we have focused only on 3 prominent LLMs, excluding other LLMs including BingAI and Perplexity. At the time of the design of this study, ChatGPT and Google Bard

Table 4. List of studies investigating the utility of ChatGPT and Google Bard across various contexts of medicine and healthcare.

First author	Year of publication	LLMs	Domain	Questions (*n*)	Reviewers (*n*)
Al-Ashwal FY[22]	2023	ChatGPT—Google Bard—Bing AI	Drug-drug interactions	225 (OE)	NA
Dhanvijay AK[23]	2023	ChatGPT—Google Bard—Bing AI	Physiology	77 (OE)	2
Seth I[24]	2023	ChatGPT—Google Bard—Bing AI	Rhinoplasty	6 (OE)	3
Koga S[25]	2023	ChatGPT—Google Bard	Neurodegenerative disorder	25 (OE)	NA
Kumari A[26]	2023	ChatGPT—Google Bard	Hematology	50 (OE)	3
Lim ZW[27]	2023	ChatGPT—Google Bard	Myopia	31 (OE)	3
Meo SA[28]	2023	ChatGPT—Google Bard	Endocrinology, diabetes, and diabetes technology	100 (MC)	—
Toyama Y[29]	2023	ChatGPT—Google Bard	Radiology	103 (MC)	3
Waisberg E[30]	2023	ChatGPT—Google Bard	Ophthalmology	NA	4
Zuniga Salazar G[31]	2023	ChatGPT—Google Bard—Bing AI	Emergency	176 (OE)	NA

Abbreviations: MC, multiple choice; NA, not available; OE, open-ended.

were the most investigated LLMs and, thus, we elected to focus on them. However, recent evidence has shown the potential of BingAI in the biomedical field.[22,37] Therefore, our results do not represent the entire spectrum of LLMs available and further assessment of other LLMs in the field of IO is essential. Secondly, the results of this study are derived only from the responses deemed "reproducible" by the reviewers. The remaining answers were not further analyzed and, thus, not considered for the final evaluation of the LLMs. Secondly, the rating process of the answers was made by only 2 reviewers. Furthermore, the limited sample size, the small number of reviewers, and the use of either Boolean or 3-point Likert scales to assess the answers (potentially resulting in the loss of subtle or nuanced differences in the responses) does limit the generalizability of these data. Thirdly, as noted above, ChatGPT was trained on specific datasets only valid up to September 2021. On the contrary, other LLMs (including BingAI) can remain up to date by continuously accessing real-time internet search data[18]—this being critical for their use in medicine and healthcare. Notably, this is also particularly relevant in the IO field due to rapidly evolving data with resultant major changes to treatment and management paradigms. Finally, the number of open-ended questions included was relatively small, which may have impacted the analysis, particularly for domain-specific performance.

Conclusion

ChatGPT-3.5 and ChatGPT-4 have demonstrated significant and clinically meaningful utility as decision- and research-aids in various subfields of IO, while Google Bard demonstrated significant limitations, especially in comparison to ChatGPT. However, the risk of inaccurate or incomplete responses was evident in all LLMs, highlighting the importance of an expert-driven verification of the information provided by these technologies. Finally, despite their potential to positively impact the field of medicine and healthcare, this study reinforced the significance of a human evaluation of LLMs in order to create reliable tools for clinical use.

Funding

This work was supported by the Intramural Research Program, National Institutes of Health, National Cancer Institute, Center for Cancer Research. The interpretation and reporting of these data are the sole responsibility of the authors.

Conflict of Interest

The authors indicated no financial relationships.

Author Contributions

Conceptualization (G.M.I., D.B.C., C.S.F.); Formal analysis (G.M.I. and H.C.W.); Investigation (G.M.I. and D.B.C.); Methodology (G.M.I. and D.B.C.); Visualization (G.M.I. and C.S.F.); Writing—Original Draft (G.M.I., D.B.C., H.C.W.); Writing—Review & Editing (F.K., J.G., C.S.F.); Supervision (C.S.F.). All authors accepted the final draft of the manuscript.

Data Availability

The data underlying this article are available in the article and in Spplementary Material.

Supplementary Material

Supplementary material is available at *The Oncologist* online.

References

1. IBM. What is generative AI?. 2021 https://research.ibm.com/blog/what-is-generative-AI
2. IBM. *What is Natural Language Processing?* IBM. https://www.ibm.com/topics/natural-language-processing
3. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. *Nat Rev Phys*. 2023;5(5):277-280. https://ora.ox.ac.uk/objects/uuid:9eac0305-0a9a-4e44-95f2-c67ee9e-ae15c.
4. 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(6):589-596. https://doi.org/10.1001/jamainternmed.2023.1838
5. Risk A, Petersen C. Health information on the internet: quality issues and international initiatives. *JAMA*. 2002;287(20):2713-2715. https://doi.org/10.1001/jama.287.20.2713
6. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel, Switzerland)*. 2023;11(6):887. https://doi.org/10.3390/healthcare11060887
7. Iannantuono GM, Bracken-Clarke D, Floudas CS, et al. Applications of large language models in cancer care: current evidence and future perspectives. *Front Oncol*. 2023;13:1268915. https://doi.org/10.3389/fonc.2023.1268915
8. Johnson DB, Nebhan CA, Moslehi JJ, Balko JM. Immune-checkpoint inhibitors: long-term implications of toxicity. *Nat Rev Clin Oncol*. 2022;19(4):254-267. https://doi.org/10.1038/s41571-022-00600-w
9. Darvin P, Toor SM, Sasidharan Nair V, Elkord E. Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp Mol Med*. 2018;50(12):1-11. https://doi.org/10.1038/s12276-018-0191-1
10. OpenAI. What is ChatGPT? https://help.openai.com/en/articles/6783457-what-is-chatgpt
11. Google. Try Bard, an AI experiment by Google. https://bard.google.com
12. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med*. 2012;22(3):276-282. https://doi.org/10.1016/j.jocd.2012.03.005
13. von Elm E, Altman DG, Egger M, et al; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-349. https://doi.org/10.1016/j.jclinepi.2007.11.008
14. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388(13):1201-1208. https://doi.org/10.1056/NEJMra2302038
15. McCarthy J. What is artificial intelligence? https://www-formal.stanford.edu/jmc/whatisai.pdf
16. IBM. AI vs. machine learning vs. deep learning vs. neural networks: what's the difference?. 2023. https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/
17. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv; 2020. http://arxiv.org/abs/2005.14165
18. Microsoft. Bing AI. https://www.microsoft.com/en-us/bing/do-more-with-ai?form=MA13KP
19. Perplexity AI. Perplexity. https://www.perplexity.ai/

20. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313. https://doi.org/10.1126/science.adg7879
21. Sullivan RJ, Weber JS. Immune-related toxicities of checkpoint inhibitors: mechanisms and mitigation strategies. *Nat Rev Drug Discov*. 2022;21(7):495-508. https://doi.org/10.1038/s41573-021-00259-5
22. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN. Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. *Drug Healthc Patient Saf*. 2023;15:137-147. https://doi.org/10.2147/DHPS.S425858
23. Dhanvijay AKD, Pinjar MJ, Dhokane N, et al. Performance of large language models (ChatGPT, Bing Search, and Google Bard) in solving case vignettes in physiology. *Cureus*. 2023;15(8):e42972. https://doi.org/10.7759/cureus.42972
24. Seth I, Lim B, Xie Y, et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. *Aest Surg J. Open Forum*. 2023;5:ojad084. https://doi.org/10.1093/asjof/ojad084
25. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol Zurich Switz*. 2023:e13207. https://doi.org/10.1111/bpa.13207
26. Kumari A, Kumari A, Singh A, et al. Large language models in hematology case solving: a comparative study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus*. 2023;15(8):e43861. https://doi.org/10.7759/cureus.43861
27. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95:104770. https://doi.org/10.1016/j.ebiom.2023.104770
28. Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The scientific knowledge of Bard and ChatGPT in endocrinology, diabetes, and diabetes technology: multiple-choice questions examination-based performance. *J Diabetes Sci Technol*. 2023:19322968231203987. https://doi.org/10.1177/19322968231203987
29. Toyama Y, Harigai A, Abe M, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Jpn J Radiol*. 2023. https://doi.org/10.1007/s11604-023-01491-2
30. Waisberg E, Ong J, Masalkhi M, et al. Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology. *Eye (Lond)*. 2023. https://doi.org/10.1038/s41433-023-02760-0
31. Zúñiga Salazar G, Zúñiga D, Vindel CL, et al. Efficacy of AI Chats to determine an emergency: a comparison between OpenAI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. *Cureus*. 2023;15(9):e45473. https://doi.org/10.7759/cureus.45473
32. Rahsepar AA, Tavakoli N, Kim GHJ, et al. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology*. 2023;307(5):e230922. https://doi.org/10.1148/radiol.230922
33. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. https://doi.org/10.1038/s41586-023-06291-2
34. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-226. https://doi.org/10.1038/d41586-023-00288-7
35. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0
36. Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform*. 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac409
37. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence ChatBots in the Antwerp University multiple choice medical license exam. medRxiv; 2023 https://doi.org/10.1101/2023.08.18.23294263. https://www.medrxiv.org/content/10.1101/2023.08.18.23294263v1