BioMedInformatics

MDPI

Article

Replies to Queries in Gynecologic Oncology by Bard, Bing and the Google Assistant

Edward J. Pavlik *, Dharani D. Ramaiah, Taylor A. Rives, Allison L. Swiecki-Sikora and Jamie M. Land

Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, Chandler Medical Center-Markey Cancer Center, University of Kentucky College of Medicine, Lexington, KY 40536-0293, USA; ddramaiah1@uky.edu (D.D.R.); taylor.rives@uky.edu (T.A.R.); allison.sikora@uky.edu (A.L.S.-S.); jamie.land@uky.edu (J.M.L.)

* Correspondence: edward.pavlik@uky.edu; Tel.: +1-(859)-323-3830

Abstract: When women receive a diagnosis of a gynecologic malignancy, they can have questions about their diagnosis or treatment that can result in voice queries to virtual assistants for more information. Recent advancement in artificial intelligence (AI) has transformed the landscape of medical information accessibility. The Google virtual assistant (VA) outperformed Siri, Alexa and Cortana in voice queries presented prior to the explosive implementation of AI in early 2023. The efforts presented here focus on determining if advances in AI in the last 12 months have improved the accuracy of Google VA responses related to gynecologic oncology. Previous questions were utilized to form a common basis for queries prior to 2023 and responses in 2024. Correct answers were obtained from the *UpToDate* medical resource. Responses related to gynecologic oncology were obtained using Google VA, as well as the generative AI chatbots Google Bard/Gemini and Microsoft Bing-Copilot. The AI narrative responses varied in length and positioning of answers within the response. Google Bard/Gemini achieved an 87.5% accuracy rate, while Microsoft Bing-Copilot reached 83.3%. In contrast, the Google VA's accuracy in audible responses improved from 18% prior to 2023 to 63% in 2024. While the accuracy of the Google VA has improved in the last year, it underperformed Google Bard/Gemini and Microsoft Bing-Copilot so there is considerable room for further improved accuracy.

Keywords: accuracy; virtual assistants; Google; Google Bard/Gemini; Microsoft Bing-Copilot; gynecologic; oncology

Citation: Pavlik, E.J.; Ramaiah, D.D.; Rives, T.A.; Swiecki-Sikora, A.L.; Land, J.M. Replies to Queries in Gynecologic Oncology by Bard, Bing and the Google Assistant. *BioMedInformatics* **2024**, *4*, 1773–1782. https://doi.org/10.3390/biomedinformatics4030097

Academic Editor: Hans Binder

Received: 13 February 2024
Revised: 20 May 2024
Accepted: 15 July 2024
Published: 24 July 2024

1. Introduction

The Internet is a well-known source of health information [1] which can be accessed by voice technology that searches the Internet, and addressed with a voice reply by a virtual assistant (VA). We recently reported on the performance of the Google VA, Siri (Apple), Cortana (Microsoft) and Alexa (Amazon) in general information queries and those specifically related to gynecologic oncology [2]. The most correct audible replies (83.3% correct) were generated by the Google VA for general queries unrelated to gynecologic oncology, as well as for those related to gynecologic oncology (18.1% correct). An explosive introduction of artificial intelligence into search engines occurred in the year that followed our report, with the release of ChatGPT-3.5 (30 November 2022), ChatGPT-4 (14 March 2023) and ChatGPT-4 Turbo (November 2023), developed by Open AI [3]. It is well-recognized that the ChatGPT family has been developed to sound coherent and not necessarily to be factually accurate [3]. For example, in the clinical setting, Chat GPT was reported to have an accuracy of 60.3% in forming an accurate initial differential diagnosis [4]. This performance is probably linked to the degree to which information is updated to be current. Note that the most recent update of the ChatGPT knowledgebase was in April 2023 [5]. Google Bard is a generative AI chatbot introduced in March 2023 that utilizes its own large

language AI model called Gemini [6]. Google Bard/Gemini pulls current information from the Internet and is available in over 40 languages. Microsoft has discontinued Cortana and launched Bing Chat, known as Copilot, within its Edge browser in February 2023. It uses Prometheus, which is its own large language model that was built on the OpenAI GPT-4 foundation [7]. Present day searches on the Google Chrome browser look distinct from Google Bard/Gemini searches; however, Chrome searches can display a button to generate an "AI-powered overview" or can return generative AI results that are visually distinct from those returned by Google Bard/Gemini. The focus of this paper was to determine the extent to which utilization of AI in the last year has improved the performance of the Google VA with regard to answers for previously examined questions specific to gynecologic oncology, and to make comparisons to the AI driving Google's Bard/Gemini and Microsoft's Bing-Copilot. These efforts are important because inaccuracies in healthcare information can lead to misinterpretation by patients which can cause them to reject or withdraw from therapies that might have proven effective. In the present paper, we determined if the accuracy of responses by the Google VA to questions related to gynecologic oncology have improved since AI implementation.

2. Materials and Methods

Google VA (version 15.4.35.29.arm64) was accessed on smartphones. Google's Bard/Gemini and Microsoft's Bing-Copilot were accessed on personal computers running Windows 10 Enterprise 64 bit (version 10.0.19045 build 19045.3930 with experience pack 1000.19053.1000.0) and Windows 11 Enterprise 64 bit (version 10.0.22631 build 22631 with experience pack 1000.22684.1000.0). Bard was accessed running the Gemini family of models, including Pro 1.0, with the Bard service rebranded as Gemini. Microsoft Edge (version 121.0.2277.110) was used with Bing-Copilot being continuously updated without specifying a version. Questions specific to gynecologic oncology were exactly as previously used and reported [2]. Each question was queried five times. Evidence-based answers to queries were obtained from *UpToDate*, a subscription-based online service used as a point-of-care decision-support resource by clinical caregivers [8]. Queried responses were evaluated as a percentage of responses that were correct and in terms of location in the narrative response. Application of the correct answer from *UpToDate* is shown in Figure 1 for a Google Bard/Gemini AI narrative query response with the correct *UpToDate* information in green font. The word counts were obtained by copying and pasting the text into Microsoft Word for the entire Google Bard/Gemini AI narrative query response (463 words, Figure 1) and for the number of words that preceded the text containing the correct information obtained from *UpToDate* (180 words, top box, Figure 1). The percentage of narrative defined the location of the correct information in the Google Bard/Gemini AI narrative query response and was calculated as 180/463 = 38.9% (Figure 1). These evaluations were applied to query responses obtained from both Google Bard/Gemini and Microsoft Bing-Copilot. The results collected here were summarized with descriptive statistics using Winstat (version 2012.1). Queries were conducted on desktop computers in the University of Kentucky IT healthcare security system and reflect limitations imposed by the security system on both Google Bard/Gemini and Microsoft Bing-Copilot. These limitations are proprietary and not known to the authors, but may have had an effect on the operations of both Google Bard/Gemini and Microsoft Bing-Copilot. The access of Google Bard/Gemini to the query source was not restricted with regard to Gemini conversations, related product usage information, information about location that the query was made from and query feedback [9]. Similarly, no efforts were made to restrict Microsoft Bing-Copilot to the query source, and Microsoft states that it does not use query data to train Copilot [10]. Circumstances at the time queries were made, or as part of an evolving evolution of either Google Bard/Gemini or Microsoft Bing-Copilot, may have had an influence on queries that were made.

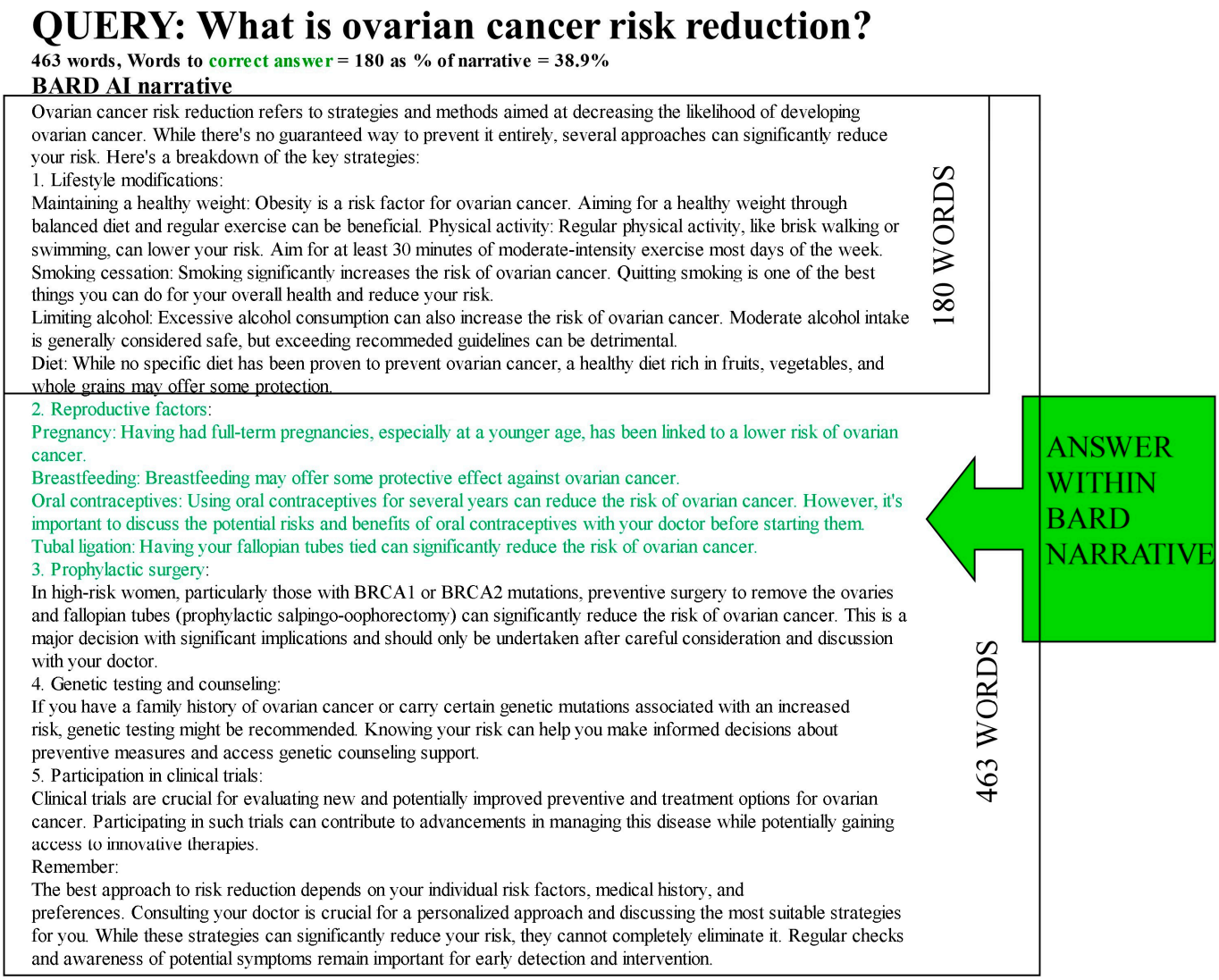

Figure 1. Determination of the position of a correct answer within the AI-generated narrative. The narrative was examined for the answer provided by *UpToDate* and marked with green font. The total narrative word count was determined (463 words = A) and word count preceding the correct answer was determined (180 words = B). The % of narrative is the position of the correct answer in the AI narrative and was determined as [(B/A) × 100].

3. Results

The questions posed as gynecologic oncology-related queries are listed in Table 1, as well as links to correct answers in *UpToDate*. In re-evaluating Google VA efforts, we first evaluated Google Bard/Gemini in order determine the degree to which the accuracy of Bard's AI narrative response was mirrored in the Google VA. Narrative length responses from Bard/Gemini changed in each repeat of a query, ranging from 32–39% of the mean narrative response length for each question, Table 1. Next, similar evaluations were made using Microsoft Bing-Copilot on the Edge browser in order to determine how similar narrative responses originating from a different large language model would be to Bard/Gemini's. The narrative response length from Microsoft Bing-Copilot also varied for each repeat of a query similarly to those from Google's Bard/Gemini. Microsoft Bing-Copilot allowed voice queries without voice recognition enabling the query, so that a mouse click on a microphone icon was needed to initiate voice recognition and supplied voice replies along with the text narrative. On Google's Bard/Gemini, a voice query could also be submitted by clicking on a microphone icon; however, the text narrative from Google's Bard/Gemini was not

accompanied with a voice reply. Thus, in terms of operation Microsoft Bing-Copilot was very similar to Google VA except for voice recognition alone being able to enable the query using the Google VA.

Table 1. Gynecologic Oncology-related Queries to Google Bard/Gemini. Narrative length returned to five repeat queries is shown as mean ± SEM with the range within parentheses. Sources of correct answers from *UpToDate* are hyperlinked for each question. All links were accessed on 22 July 2024.

Query #	Question	Correct Answer Link
1.	What is stage I ovarian cancer? [308.6 ± 15.6 (247,330)]	Answer
2.	What is stage II ovarian cancer? [349.6 ± 15.5 (310,394)]	Answer
3.	What is stage III ovarian cancer? [344.6 ± 20 (307,421)]	Answer
4.	What is stage IV ovarian cancer? [338.6 ± 27.7 (252,425)]	Answer
5.	What is stage IC1 ovarian cancer? [309.6 ± 22.1 (223,347)]	Answer
6.	What is stage IIIA1 ovarian cancer? [355 ± 16.2 (325,414)]	Answer
7.	What is stage IVB ovarian cancer? [371.8 ± 25.1 (303,436)]	Answer
8.	What are the subtypes of epithelial ovarian cancer? [296.8 ± 15.7 (365,350)]	Answer
9.	What is screening for ovarian cancer? [324 ± 23.2 (264,384)]	Answer
10.	What are the screening recommendations for ovarian cancer? [318.4 ± 15.3 (284,359)]	Answer
11.	What are ways to prevent ovarian cancer? [387.8 ± 23 (318,439)]	Answer
12.	What are the symptoms of ovarian cancer? [313.8 ± 22.9(237,365)]	Answer
13.	What is hereditary ovarian cancer? [320 ± 24.9 (256,386)]	Answer
14.	What is ovarian cancer risk reduction? [372.2 ± 43.1 (252,518)]	Answer
15.	What is screening for cervical cancer? [335.4 ± 38.2 (236,423)]	Answer
16.	What are the screening recommendations for cervical cancer? [260.8 ± 15.2 (219,350)]	Answer
17.	What are the options for a 20-year-old woman requesting a Pap smear? [383 ± 23.8 (334,473)]	Answer
18.	What is the HPV vaccine? [355.8 ± 19.2 (301,412)]	Answer
19.	What are the ages for HPV vaccination? [229.6 ± 21.4 (186,304)]	Answer
20.	What are the three dose HPV vaccine recommendations? [258.2 ± 24.8 (197,344)]	Answer
21.	What are borderline epithelial tumors of the ovary? [349.6 ± 21.9 (294,413)]	Answer
22.	What is carcinosarcoma of the ovary? [322.6 ± 7.4 (303,344)]	Answer
23.	What are high-grade serous tumors of the ovary? [311.4 ± 15.5 (277,368)]	Answer
24.	What is stage IB endometrial cancer? [338.4 ± 13.4 (295,366)]	Answer

Next, Google Bard/Gemini and Microsoft Bing-Copilot were compared directly. Their performances were similar, although Google's Bard/Gemini had a higher percentage of correct responses (87.5% Bard/Gemini, 83.3% Microsoft Bing-Copilot) (Table 2). Out of five repeated queries for each of the 24 questions, there were a small number of instances where a correct response disappeared from the narrative after being present in the preceding narratives for that query. Accordingly, if all five queries failed to generate a narrative with a correct response, then the response was assigned as "Correct = NO" for that question. Estimating correct response percentages on the basis of 120 possible responses to the 24 questions that were queried five times, Google's Bard/Gemini had a higher percentage of correct responses (82.5% Bard/Gemini, 78.3% Microsoft Bing-Copilot) (Table 2).

Using the approach presented in Figure 1, the position of a correct answer within each AI-generated narrative was determined and summarized as lying in the 0–10% frequency range, 11–25% frequency range or 26–50% frequency range of the total word count for the narrative response of either Google Bard/Gemini or Microsoft Bing-Copilot (Table 2). Responses that occurred early (0–10% frequency range), later (11–25% frequency range) or even later (26–50% frequency range) in the word count of narrative responses were not significantly different (p = 0.496) between Google Bard/Gemini (66.7%, 23.8% and 9.5%)

and Microsoft Bing-Copilot (50%, 30% and 20%), as determined by Chi-square analysis in a contingency table of the number of responses in each frequency distribution range for Google Bard/Gemini vs Microsoft Bing-Copilot. Thus, Google Bard/Gemini and Microsoft Bing-Copilot performed similarly in accuracy (% correct responses) and positioning the correct information within their respective AI-generated narratives.

Table 2. Responses to Gynecologic Oncology-related Questions by Google Bard/Gemini via the Chrome browser and Microsoft Bing-Copilot via the Edge browser. Convention for reporting of responses for each question: correctly answered "YES" or "NO" (word distance to correct answer as a percentage of entire narrative response). Questions were scored as correctly answered if any of the five queries for any question were correct. Separate tabulations of correct queries, shown in parentheses on the last line, were based on 120 queries for all 24 questions.

Question:		Google-Bard/Gemini Correct (YES/NO)	MS-Bing-CoPilot Correct (YES/NO)
1.	What is stage I ovarian cancer?	YES (3.3%)	YES (5.3%)
2.	What is stage II ovarian cancer?	YES (1.4%)	YES (0.4%)
3.	What is stage III ovarian cancer?	YES (8.7%)	YES (14.5%)
4.	What is stage IV ovarian cancer?	YES (4.2%)	YES (4.2%)
5.	What is stage IC1 ovarian cancer?	YES (28%)	YES (25.6%)
6.	What is stage IIIA1 ovarian cancer?	YES (9.1%)	YES (19%)
7.	What is stage IVB ovarian cancer?	YES (7.1%)	YES (11.6%)
8.	What are the subtypes of epithelial ovarian cancer?	YES (8.4%)	YES (27.6%)
9.	What is screening for ovarian cancer?	YES (9.9%)	YES (16%)
10.	What are the screening recommendations for ovarian cancer?	YES (0.3%)	YES (0.5%)
11.	What are ways to prevent ovarian cancer?	YES (32%)	YES (24.4%)
12.	What are the symptoms of ovarian cancer?	YES (12.8%)	YES (25.9%)
13.	What is hereditary ovarian cancer?	YES (4.6%)	YES (8.2%)
14.	What is ovarian cancer risk reduction?	YES (22%)	YES (28.3%)
15.	What is screening for cervical cancer?	YES (18.2%)	YES (13.7%)
16.	What are the screening recommendations for cervical cancer?	NO	NO
17.	What are the options for a 20-year-old woman requesting a Pap smear?	YES (9.7%)	YES (10.2%)
18.	What is the HPV vaccine?	YES (0.3%)	YES (0.5%)
19.	What are the ages for HPV vaccination?	NO	NO

Table 2. *Cont.*

Question:	Google-Bard/Gemini Correct (YES/NO)	MS-Bing-CoPilot Correct (YES/NO)
20. What are the three dose HPV vaccine recommendations?	NO	NO
21. What are borderline epithelial tumors of the ovary?	YES (3.9%)	YES (17.9%)
22. What is carcinosarcoma of the ovary?	YES (13.5%)	NO
23. What are high-grade serous tumors of the ovary?	YES (17.3%)	YES (10.9%)
24. What is stage IB endometrial cancer?	YES (13.7%)	YES (0.5%)
Total number of correct responses	21	20
Percentage of correct number of questions (% correct queries)	87.5% (82.5%)	83.3% (78.3%)

Finally, when Google VA was re-queried in 2024 with the same questions that were used before 2023, where an accuracy rate of 18% correct auditory responses was found [2], performance increased to 63% (Figure 2). While the accuracy of the Google VA has improved, it remains lower than the accuracy of audible replies to queries in gynecologic oncology returned by Google Bard/Gemini and Microsoft Bing-Copilot.

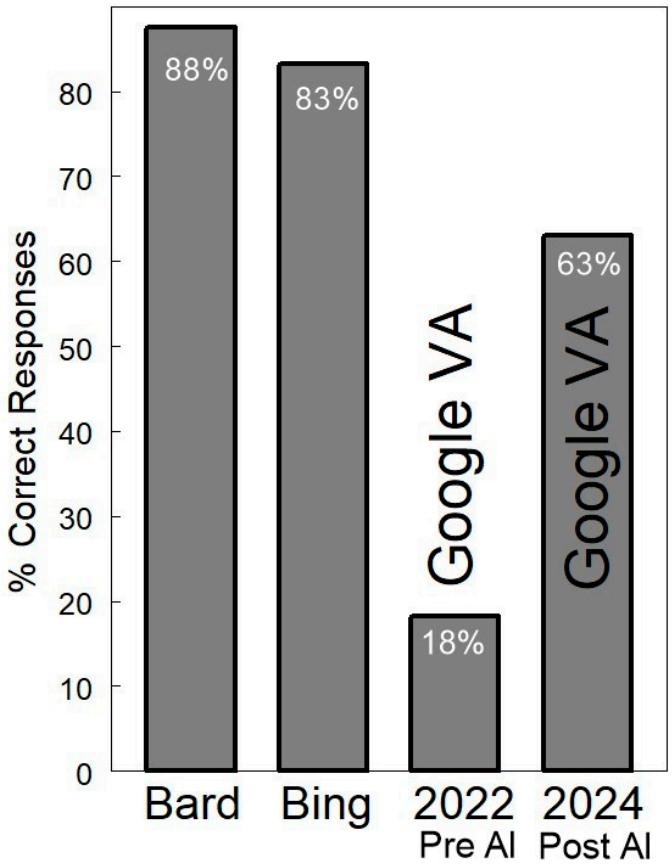

Figure 2. Comparative Performance of Google Bard/Gemini, Microsoft Bing-Copilot and Google virtual assistant (VA) in 2022 and 2024.

4. Discussion

In summary, replies by the Google VA to gynecologic oncology-related queries have improved considerably in the last year, but are not as accurate as narrative replies from Google Bard/Gemini or Microsoft Bing-Copilot. In the year following the announcements on narrative AIs heralding their various advantages and improvements, reports have also been published questioning chatbot accuracy. ChatGPT-3.5$_{\text{turbo-0301}}$ performed poorly in providing accurate cancer treatment recommendations, and generated outputs that were not concordant with NCCN recommendations a third of the time [11]. In answering "Questions to Ask About Your Cancer", recommended by the American Cancer Society, both ChatGPT-3.5 and Bing answered correctly less than 80% of the time [12]. ChatGPT-4 was found to be capable of correctly diagnosing only 57% of complex clinical cases [13]. However, on a board-style neurology examination ChatGPT4 was able to answer 85% correct on over 1900 questions in behavioral, perceptive and psychological-related areas, using confident language for both correct as well as incorrect answers [14]. Furthermore, when asked to provide the energy content of 222 food items, both ChatGPT-3.5 and ChatGPT-4 provided accurate answers less than half of the time [15]. From experiences related to the analysis presented in this paper, several characteristics of the narrative responses should be noted. No narrative included a reference to *UpToDate*, suggesting that sources of information behind pay walls are excluded from information gathering by Google Bard/Gemini and Microsoft Bing-Copilot. The degree to which this exclusion applies to other narrative chat AI is an open question. From a clinical standpoint, because *UpToDate* has long-standing universal acceptability as a point-of-care decision support resource that is very frequently updated, it is difficult to expect acceptability of narratives that exclude *UpToDate* from their narrative responses to queries, no matter how confident, well-presented and high quality the narrative text is. It should be pointed out that Google Bard/Gemini does qualify their narrative results: "This is for informational purposes only. This information does not constitute medical advice or diagnosis". However, the accuracy of the information is not perfect. Microsoft Bing-Copilot's narrative replies: "It is important to consult with healthcare professionals for accurate information". This statement leaves unanswered the question as to whether healthcare professionals should look elsewhere than Microsoft Bing-Copilot for their information. There are several situations that are pertinent to the work presented here. First, some of the questions that we have utilized may occur to a family member of a patient seen in gynecologic oncology. Such a narrative AI inquiry certainly should be considered with the possibility that it may not be as accurate as information conveyed by a gynecologic oncology specialist. It is important that inaccurate information and the family member's advice do not interfere with a patient's decision to continue treatment. Second, a gynecologic oncology patient may need to further their understanding of something that their physician/specialist conveyed to them. In this context, their gynecologic oncology specialist should be aware that their patient may be subject to influence by information the patient has received from a narrative AI that is less accurate than information that has been related to them by the physician. Third, because ovarian cancer treatment is determined by stage and new treatments appear several times a year, a generalist physician in an emergency room setting may need to clarify factors affecting a patient's underlying health. An understanding of the accuracy of responses obtainable through narrative AI should be considered and weighed against inquiries on *UpToDate* and the National Comprehensive Cancer Network (nccn.org) as sources more likely to relate treatment strategies and how they may affect a patient's feeling of well-being. In a similar sense, mid-level providers at a walk-in treatment center should also understand the most accurate sources of specialty information that are relevant to complaints by a newly arrived patient. The significance of the work presented here informs readers that some inaccuracy can exist in narrative AI information and should be balanced against other sources. It should be clearly understood that our report relates to the free Google VA, free Google Bard/Gemini and free Microsoft Bing-Copilot. We do not make any extension of the conclusions raised here by inference or direct examination to other AI-based platforms.

The strengths of the present paper include: (1) use of the same 24 question set that was previously evaluated in order to make precise comparisons to current performance by the Google VA, Google Bard/Gemini and Microsoft Bing Copilot, (2) utilization of a continuously updated point-of-care decision-support resource compiled from more than 440 journals by clinicians as the source of correct answers [8], (3) evaluations of the Google VA in comparison to Google Bard/Gemini and Microsoft Bing Copilot, (4) multiple examinations of the length of each response to a query, (5) identification of how deeply imbedded the correct response was within a narrative response and (6) use of speech-to-text conversion to find and validate correctness of answers from the Google VA. The limitations of this work involve: (1) inability to determine how future-stable responses by the Google VA, Google Bard/Gemini and Microsoft Bing Copilot are for either accuracy or inaccuracy, (2) exclusion of efforts to broadly assess the findings here to parallel queries in clinical areas other than gynecologic oncology, (3) exclusion of efforts to determine if and when voice activation will be active for Google Bard/Gemini and Microsoft Bing Copilot so that further improved accuracy will become available, (4) limitations undefined to the authors that are imposed by the institutional security system and (5) advantages or disadvantages arising from access or non-access by Google Bard/Gemini or Microsoft Bing-Copilot to data on the desktop computers that originated the queries. The voice-activated VA will be most valuable when users cannot manually enter information, and this will be very helpful to the elderly, to physically-impaired individuals and to those that are bed-ridden. In situations where a resource like *UpToDate* cannot be utilized to evaluate the accuracy of AI responses, a degree of risk in the trustworthiness of responses to queries must be anticipated. For example, in queries about drug cross reactions or about poisons, the consequences of inaccuracies can be deadly. There are multiple ways that can provide users with greater confidence about the accuracy of AI search results. The use of a "custom attribution engine" as announced by Adobe would allow users to verify AI findings through source citation [16]. This type of approach should allow users to interpret narrative results in terms of source information and determine if there is any distortion in the AI narrative. Importantly, for narratives that contain clinical information, it is important to gauge the results against information resources that are frequently updated through efforts involving a large number of medical experts, like *UpToDate* or the National Comprehensive Cancer Network (nccn.org). Improvement in these AI platforms will occur when benchmarking the accuracy of results becomes an intrinsic process that is motivated by external evaluations as presented here. Ideally, accuracy rating scores should be created to accompany individual narrative response. Each AI narrative response should be made to provide reference lists identifying the source of material and if re-uses of AI-generated information retrievals were employed so that users are provided with a transparent view of the sources of information retrieved.

Different semantic interpretations of the performance of narrative AI do exist. For example, some focus can be on relevance, comprehensiveness and clarity of the information provided to the users. In the approach that we have taken, relevance could relate to both the question posed and the answers returned by the Google VA, Google Bard/Gemini and Microsoft Bing-Copilot. Relevance in terms of the queries was related only to their prior utilization [2]. Relevance in terms of the responses returned by the Google VA, Google Bard/Gemini and Microsoft Bing-Copilot was determined by the synchrony between the context of the query and that of the information in the response. As an example, consider Query 4, *What is stage IV ovarian cancer?* The responses returned by the Google VA, Google Bard/Gemini and Microsoft Bing-Copilot would not be relevant if the stage was stated as III or if uterine cancer was stated in the reply. Comprehensiveness, as the state or condition of including all or nearly all elements or aspects of something, is not the expectation for a virtual assistant like the Google VA. Comprehensiveness for responses to queries on Google Bard/Gemini and Microsoft Bing-Copilot is a subjective concept. An individual making a query probably has an expectation that an answer is not buried in an extensive tome of retrieved information. We have presented word counts as a measure of extensiveness in

replies by Google Bard/Gemini and Microsoft Bing-Copilot. We have also included word distance to the correct answer within the narrative constructed by Google Bard/Gemini and Microsoft Bing-Copilot. Taken together, our report indicates that the narratives returned by Google Bard/Gemini and Microsoft Bing-Copilot offer more than a succinct correct answer; however, addressing the degree to which the narratives are comprehensive is not something that was addressed. Moreover, determining the degree to which all aspects of factors identified in the query are included in narrative responses might not be the intent of those implementing the narrative responses. Clarity of narrative responses also is a highly subjective determination. Clarity in writing refers to being clear and concise to your intended audience. Clear writing communicates ideas effectively, without any ambiguity or confusion. It involves using plain language and avoiding jargon that might be unfamiliar to the reader. Clarity in the context of the narratives returned by the Google VA, Google Bard/Gemini and Microsoft Bing-Copilot was only determined by evaluators being able to identify the correct answer.

Taken together, there is ample evidence that echoes our findings that there is room for improved accuracy in the Google VA, Google Bard/Gemini and Microsoft Bing Copilot. Inaccuracy can pose a danger in medical settings as it can be stated in a manner in which the VA or the AI narrative is very confident and convincing. While a recent editorial acknowledges that, "for use in publication, large language models present concerns regarding authorship, originality, factual inaccuracies and "hallucinations" or confabulations, the stated key to their acceptability is that authors take complete responsibility for the content and properly acknowledge the use of LLMs" [17]. In short, the accuracy of a clinically-related publication remains the responsibility of the authors. As reported recently, an airline was unsuccessful in establishing that it was not responsible on its own corporate site for information provided by a chatbot [18]. Thus, the precedent has been set for legal responsibility and liability related to chatbot utilization.

5. Conclusions

This paper underscores the need for improvements in the accuracy of information related to gynecologic oncology supplied by the Google VA, Google Bard/Gemini and Microsoft Bing-Copilot as considered here. Although it has been reported that Google's Bard/Gemini has been found to provide erroneous information related to discoveries made by the James Webb Space telescope [19,20], while Microsoft's AI-powered Bing-Copilot has also been susceptible to providing false information [21,22], the work that is presented here updates the state of accuracy and reliability to early 2024.

We conclude that audible replies by the Google VA to gynecologic oncology-related voice queries still have appreciable room for improved accuracy. Overall, we advise that patients exercise caution in the use of VAs that provide information in gynecologic oncology.

Author Contributions: Conceptualization, E.J.P. and J.M.L.; methodology, J.M.L.; software, E.J.P. and J.M.L.; validation, E.J.P.; formal analysis, E.J.P.; investigation, J.M.L., D.D.R., T.A.R. and A.L.S.-S.; resources, E.J.P.; data curation, E.J.P.; writing—original draft preparation, E.J.P.; writing—review and editing, J.M.L., D.D.R., T.A.R., A.L.S.-S. and E.J.P.; visualization, J.M.L.; supervision, E.J.P.; project administration, E.J.P.; funding acquisition, E.J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not have external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data will be made available upon request.

Conflicts of Interest: The authors declare no conflicts of interest. Text in this paper was NOT generated using any large language model.

References

1. Fox, S.; Duggan, M. Health Online 2013. Pew Research Center 2013. Available online: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/ (accessed on 22 July 2024).
2. Land, J.M.; Pavlik, E.J.; Ueland, E.; Ueland, S.; Per, N.; Quick, K.; Gorski, J.W.; Riggs, M.J.; Hutchcraft, M.L.; Llanora, J.D.; et al. Evaluation of Replies to Voice Queries in Gynecologic Oncology by Virtual Assistants Siri, Alexa, Google, and Cortana. *BioMedInformatics* **2023**, *3*, 553–562. [CrossRef]
3. Brandl, R.; Ellis, C. Tooltester. ChatGPT Statistics 2024—All the Latest Statistics about OpenAI's Chatbot. Available online: https://www.tooltester.com/en/blog/chatgpt-statistics/ (accessed on 22 July 2024).
4. Rao, A.; Pang, M.; Kim, J.; Kaminen, M.; Lie, W.; Prasad, A.K.; Landman, A.; Dreyer, K.; Succi, M.D. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J. Med. Internet Res.* **2023**, *25*, e48659. [CrossRef] [PubMed]
5. Kan, M. ChatGPT's Knowledge Base Finally Extends Beyond 2021. PC 11/06/2023. Available online: https://www.pcmag.com/news/chatgpts-knowledge-base-finally-extends-beyond-2021 (accessed on 22 July 2024).
6. Ortiz, S. ZDNET What Is Google Bard? Here's Everything You Need to Know. 9 February 2024. Available online: https://www.zdnet.com/article/what-is-google-bard-heres-everything-you-need-to-know/ (accessed on 22 July 2024).
7. Microsoft Copilot. Wikipedia. Available online: https://en.wikipedia.org/wiki/Microsoft_Copilot (accessed on 22 July 2024).
8. UpToDate. Wikipedia. Available online: https://en.wikipedia.org/wiki/UpToDate (accessed on 22 July 2024).
9. Gemini Apps Privacy Notice. Your Data and Gemini Apps. Last Updated: 8 February 2024. Available online: https://support.google.com/gemini/answer/13594961?visit_id=638450180634226006-3120607282&p=privacy_help&rd=1&collected_data#your_data (accessed on 22 July 2024).
10. What Happens to My Data When I Use Copilot? Available online: https://learn.microsoft.com/en-us/power-platform/faqs-copilot-data-security-privacy (accessed on 22 July 2024).
11. Chen, S.; Kann, B.H.; Foote, M.B.; Aerts, H.J.W.L.; Savova, G.K.; Mak, R.H.; Bitterman, D.S. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol.* **2023**, *9*, 1459–1462. [CrossRef] [PubMed]
12. Janopaul-Naylor, J.R.; Koo, A.; Qian, D.C.; McCall, N.S.; Liu, Y.; Patel, S.A. Physician Assessment of ChatGPT and Bing Answers to American Cancer Society's Questions to Ask About Your Cancer. *Am. J. Clin. Oncol.* **2024**, *47*, 17–21. [CrossRef] [PubMed]
13. Shea, Y.F.; Lee, C.M.Y.; Ip, W.C.T.; Luk, D.W.A.; Wong, S.S.W. Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis. *JAMA Netw. Open.* **2023**, *6*, e2325000. [CrossRef] [PubMed]
14. Schubert, M.C.; Wick, W.; Venkataramani, V. Performance of Large Language Models on a Neurology Board-Style Examination. *JAMA Netw. Open.* **2023**, *6*, e2346721. [CrossRef] [PubMed]
15. Hoang, Y.N.; Chen, Y.L.; Ho, D.K.N.; Chiu, W.C.; Cheah, K.J.; Mayasari, N.R.; Chang, J.S. Consistency and Accuracy of Artificial Intelligence for Providing Nutritional Information. *JAMA Netw. Open.* **2023**, *6*, e2350367. [CrossRef] [PubMed]
16. Irwin, K. Adobe's AI Assistant Can Summarize PDFs, PowerPoints for You. PC Magazine 02. 20 February 2024. Available online: https://www.pcmag.com/news/adobes-ai-assistant-can-summarize-pdfs-powerpoints-for-you?utm_source=email&utm_campaign=whatsnewnow&zdee=gAAAAABlzjXaj__BrV5eo-f2qF7sDrWWKEtkmH2G19SKAH7kdCyO2QZWMpWIrsyK4J9MUmTylA0dKjizA-4gORPeq8yq9vplTV5ka-cM6LR2CdQFveA56U4= (accessed on 22 July 2024).
17. Koller, D.; Beam, A.; Manrai, A.; Ashley, E.; Liu, X.; Gichoya, J.; Holmes, C.; Zou, J.; Dagan, N.; Wong, T.Y.; et al. Why We Support and Encourage the Use of Large Language Models in NEJM AI Submissions. *NEJM AI* **2023**, *1*. [CrossRef]
18. Cecco, L. Air Canada Ordered to Pay Customer Who Was Misled by Airline's Chatbot. The Guardian. 21 February 2024. Available online: https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit (accessed on 22 July 2024).
19. Miao, H. Alphabet Stock Drops 8% after Google Rollout of AI Search Features. The Wall Street Journal. Available online: https://www.wsj.com/livecoverage/stock-market-news-today-02-08-2023/card/alphabet-stock-drops-after-google-parent-introduces-ai-search-features-wgCJG3IDoSbfL3SgyrNI (accessed on 22 July 2024).
20. Mihalcik, C. Google ChatGPT Rival Bard Flubs Fact About NASA's Webb Space Telescope. Available online: https://www.cnet.com/science/space/googles-chatgpt-rival-bard-called-out-for-nasa-webb-space-telescope-error/ (accessed on 22 July 2024).
21. Hao, K. What Is ChatGPT? What to Know About the AI Chatbot That Will Power Microsoft Bing. The Wall Street Journal. 16 May 2023. Available online: https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177?st=q4wbp2 (accessed on 22 July 2024).
22. Quach, K. Microsoft's AI Bing Also Factually Wrong, Fabricated Text During Launch Demo. The Register. Available online: https://www.theregister.com/2023/02/14/microsoft_ai_bing_error/ (accessed on 22 July 2023).