SIUJ Société Internationale d'Urologie Journal

Article

Responses of Artificial Intelligence Chatbots to Testosterone Replacement Therapy: Patients Beware!

Herleen Pabla [1], Alyssa Lange [2], Nagalakshmi Nadiminty [1] and Puneet Sindhwani [1,*]

[1] Department of Urology, University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, USA; herleen.pabla@utoledo.edu (H.P.); nagalakshmi.nadiminty@utoledo.edu (N.N.)
[2] University of Toledo Medical School, University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, USA; alyssa.lange2@rockets.utoledo.edu
* Correspondence: puneet.sindhwani@utoledo.edu

Abstract: Background/Objectives: Using chatbots to seek healthcare information is becoming more popular. Misinformation and gaps in knowledge exist regarding the risk and benefits of testosterone replacement therapy (TRT). We aimed to assess and compare the quality and readability of responses generated by four AI chatbots. **Methods**: ChatGPT, Google Bard, Bing Chat, and Perplexity AI were asked the same eleven questions regarding TRT. The responses were evaluated by four reviewers using DISCERN and Patient Education Materials Assessment Tool (PEMAT) questionnaires. Readability was assessed using the Readability Scoring system v2.0. to calculate the Flesch–Kincaid Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL). Kruskal–Wallis statistics were completed using GraphPad Prism V10.1.0. **Results**: Google Bard received the highest DISCERN (56.5) and PEMAT (96% understandability and 74% actionability), demonstrating the highest quality. The readability scores ranged from eleventh-grade level to college level, with Perplexity outperforming the other chatbots. Significant differences were found in understandability between Bing and Google Bard, DISCERN scores between Bing and Google Bard, FRES between ChatGPT and Perplexity, and FKGL scoring between ChatGPT and Perplexity AI. **Conclusions**: ChatGPT and Google Bard were the top performers based on their quality, understandability, and actionability. Despite Perplexity scoring higher in readability, the generated text still maintained an eleventh-grade complexity. Perplexity stood out for its extensive use of citations; however, it offered repetitive answers despite the diversity of questions posed to it. Google Bard demonstrated a high level of detail in its answers, offering additional value through visual aids. With improvements in technology, these AI chatbots may improve. Until then, patients and providers should be aware of the strengths and shortcomings of each.

Keywords: artificial intelligence; misinformation; testosterone

Academic Editor: Peter C. Black

Received: 18 September 2024
Revised: 7 December 2024
Accepted: 14 January 2025
Published: 12 February 2025

Citation: Pabla, H.; Lange, A.; Nadiminty, N.; Sindhwani, P. Responses of Artificial Intelligence Chatbots to Testosterone Replacement Therapy: Patients Beware! *Soc. Int. Urol. J.* **2025**, *6*, 13. https://doi.org/10.3390/siuj6010013

1. Introduction

Patients have access to a diverse range of resources to seek healthcare information, from consulting a healthcare professional to gathering information from family and friends, print media, broadcast media, and the internet. Although healthcare providers are the most trusted sources, people turn to the World Wide Web frequently before consulting their physicians [1]. A survey on the topic showed that 80% of US adults searched online for information about a range of health issues [2]. However, online health information lacks proper regulation, allowing unrestricted contributions from individuals, compromising its reliability [3]. Recently, there has been a rising trend in utilizing chatbots like ChatGPT to

acquire healthcare information. They are based on a novel AI technology known as large language models (LLMs), which are trained on a vast amount of information from across the web, including books, articles, and websites, to generate human-like text [4]. They utilize two domains—natural language processing and deep learning—to answer prompts generated by users [5,6]. There has been a significant improvement in the development and use of chatbots, with applications in various domains, including healthcare [7]. These AI-powered chatbots are easy to access, available around the clock, and use a chat-based interactive model that allows patients to consult them to obtain information on a wide range of topics within seconds. Along with the advancements, it has been discovered that these AI models often generate false or misleading information and present it as a fact, a phenomenon popularly known as the "hallucination effect" [8,9]. Alkaissi and McFarlane discuss this phenomenon in detail, highlighting an instance where ChatGPT provided inaccurate information on the mechanism of homocysteine-induced osteoporosis and, when prompted, provided incorrect citations [10].

Testosterone replacement therapy (TRT) is a subject clouded by misinformation, and there are existing gaps in knowledge about the benefits and risks among patients, with a survey showing 50% of respondents being unaware of the risks. The internet was found to be one of the top sources for accessing information on TRT [11]. The symptoms of low testosterone, including reduced energy, sex drive, and erectile function, can have associated stigma, leading to patients resorting to online resources to address their concerns. With the advent of multiple AI chatbots and their increasing use, it becomes imperative to further analyze the quality, accuracy, and readability of the information generated by them. While several studies have assessed the quality and readability of responses generated by AI chatbots on diverse health topics, to the best of our knowledge, no literature has analyzed AI chatbot responses related to testosterone replacement therapy. Therefore, the objective of our study was to assess and compare the accuracy, quality, and readability of responses generated by four popular AI chatbots—Google Bard, ChatGPT3.5, Perplexity AI, and Bing Chat—concerning TRT.

2. Materials and Methods

2.1. Selection of Chatbots

Four AI chatbots—ChatGPT version 3.5, Google Bard (rebranded as Gemini), Bing Chat (rebranded as Copilot), and Perplexity AI—were chosen in order to answer freely formulated and complex questions. The chatbots were selected based on feedback from urology attendings and residents who identified those which were popular among their patient population. The latest free versions of these chatbots were selected, as they are readily available and widely accessible by patients.

2.2. Question Source

The questions were prepared after deliberation between a board-certified urologist and medical students based on patient experiences and commonly asked questions about TRT encountered in the outpatient clinic (Figure 1). The questions were asked in simple language to simulate a patient's perspective. To prevent bias, the chatbots were accessed in incognito mode to prevent browser history from affecting the answers. The questions were asked only once, in a sequential order, and remained the same across all chatbot engines. The answers generated by all the chatbots are provided in Supplementary Data S1.

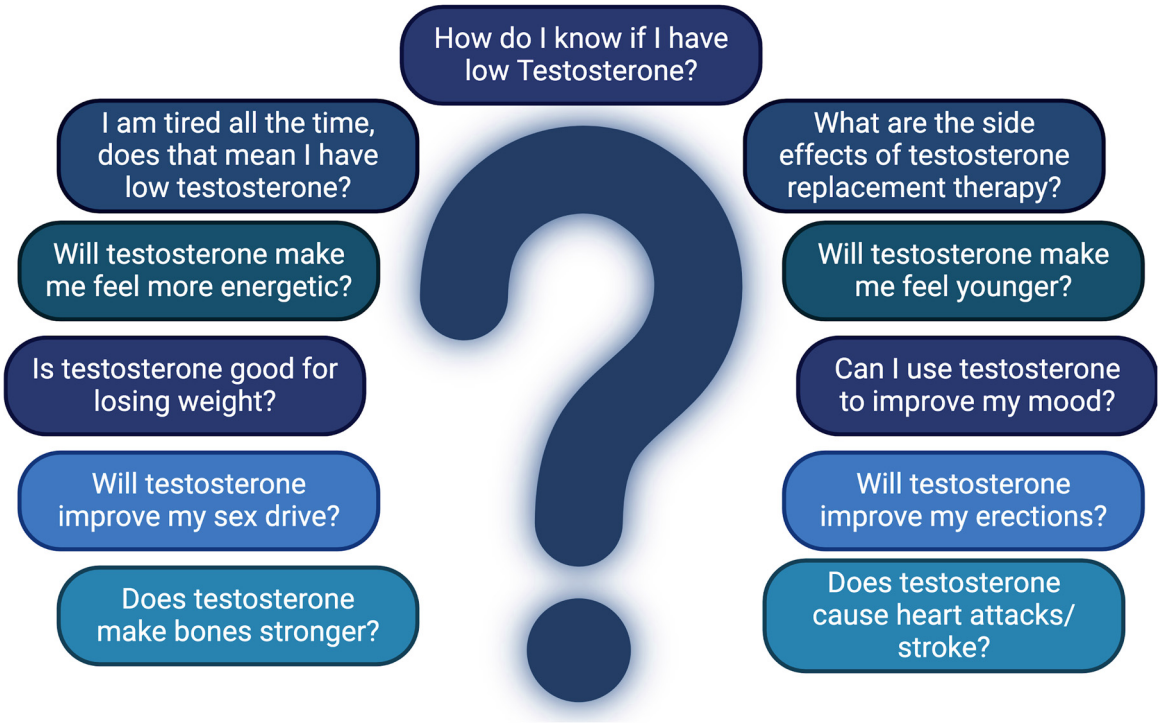

Figure 1. The questions posed to each chatbot simulating a patient's perspective.

2.3. Quality and Readability Analysis

To analyze the accuracy, the responses were compared to published and peer-reviewed guidelines from the AUA. The responses were evaluated by four independent reviewers—one board-certified urologist, one PhD urologist, and MS III and IV. The primary outcomes were the quality of consumer health information based on the validated DISCERN instrument [12] and the understandability and actionability via the Patient Education Materials Assessment Tool (PEMAT) [13]. DISCERN employs a score ranging from 1 = low to 5 = high for overall information quality and a set of 15 questions, giving a total score ranging from 15 to 75. Scores between 75 and 63 points are classified as "excellent", between 62 and 51 points as "good", between 50 and 39 as "average", between 38 and 28 as "poor", and <27 points as "very poor" [14]. PEMAT scores range between 0 and 100%, with higher scores indicating a higher level of understandability and actionability [13].

Readability was assessed using the Flesch–Kincaid Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL), using an online readability calculator [15]. The FRES ranges from 0 to 100, with a higher score corresponding to a text which is easier to read (recommended score ≥ 70), while the FKGL measures the grade level necessary to comprehend the text (recommended score of ≤ 6). An example question and answer generated by ChatGPT is given in Figure 2.

2.4. Statistical Analysis

The Kruskal-Wallis statistics were obtained using GraphPad Prism V10.1.0.

2.5. Ethical Clearance

Since the study did not involve human subjects, IRB approval was not required.

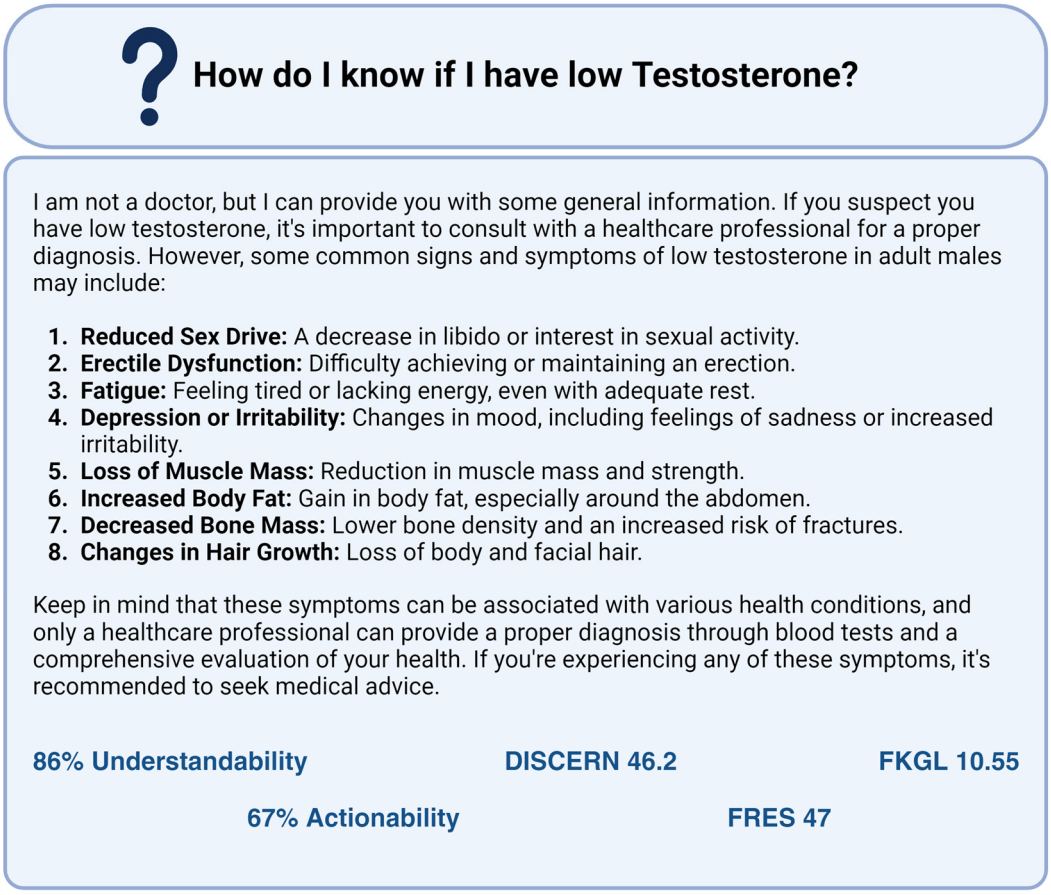

Figure 2. ChatGPT's response to one of the questions and a summary of the scores.

3. Results

The analysis included responses to eleven questions asked in sequential order, graded by four independent reviewers. The median score of each response was taken.

3.1. Quality of Information

The DISCERN scores were evaluated. The responses generated by Bing Chat were given a median score of 40 across all four reviewers, followed by ChatGPT with 46.2, Perplexity AI with 48.5, and Google Bard with the highest score of 56.5.

3.2. Understandability and Actionability of Information

For the PEMAT understandability scores, Bing AI had the lowest median (IQR) at 57% (50–60%), Perplexity AI received 74% (66–73%), ChatGPT had 86% (83–91%), and Google Bard had the highest at 96% (86–100%). The PEMAT actionability scores for Bing Chat and Perplexity were the lowest, with 40% (20–60%) and 40% (40–40%), respectively. ChatGPT received a score of 67% (60–80%), and Google Bard again had the highest score of 74% (60–83.3%).

3.3. Readability

The readability of the four AI responses was graded by the Flesch–Kincaid Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL). Perplexity AI had the highest median score of 41.9 and a range of (28–69), Bing Chat received 39.3 (27–62), Google Bard's score was 32.1 (16–52), and ChatGPT had the lowest score at 25.1 (11–47). These readability scores were all under the recommended score of 70. Perplexity AI had the best FKGL score of

10.8, followed by Bing Chat at 12.3, Google Bard at 12.7, and the lowest score was for ChatGPT at 14.9. Table 1 shows a summary of the scores received by each of the four AI chatbots.

Table 1. Summary of median (IQR) scores of each of the four AI chatbots.

TOOL	AI CHATBOT			
	Bing Chat	**ChatGPT**	**Google Bard**	**Perplexity AI**
DISCERN [1]	40 (38–44)	46.2 (43–49)	56.5 (54–58)	48.5 (42–53)
PEMAT Understandability [2]	57% (50–60%)	86% (83–91%)	96% (86–100%)	74% (66–73%)
PEMAT Actionability [2]	40% (20–60%)	67% (60–80%)	74% (60–83.3%)	40% (40–40%)
FRES [3]	39.3 (27–62)	25.1 (11–47)	32.1 (16–52)	41.9 (28–69)
FKGL [4]	12.3 (7–14.9)	14.9 (10.5–17.6)	12.7 (9.6–16.2)	10.8 (5.6–14.7)

PEMAT, Patient Education Material Assessment Tool; FRES, Flesch–Kincaid Reading Ease Score; a and FKGL, Flesch–Kincaid Grade Level. [1] Score ranging from 15 (poor) to 75 (excellent), [2] score ranging from 0% (low) to 100% (high), [3] score ranging from 0 (difficult to read) to 100 (easy to read), and [4] score measuring grade level, with a higher score corresponding to a text which is more difficult to read. The background colors represent performance, with green indicating the best performance, and red the lowest.

To assess differences in understandability, actionability, and readability, the Kruskal–Wallis test was performed using GraphPad Prism V10.1.0 with Alpha set to 0.05. See Figure 3. There were significant differences in the understandability and DISCERN scores between Bing and Google Bard and in the FRES and FKGL readability between ChatGPT and Perplexity.

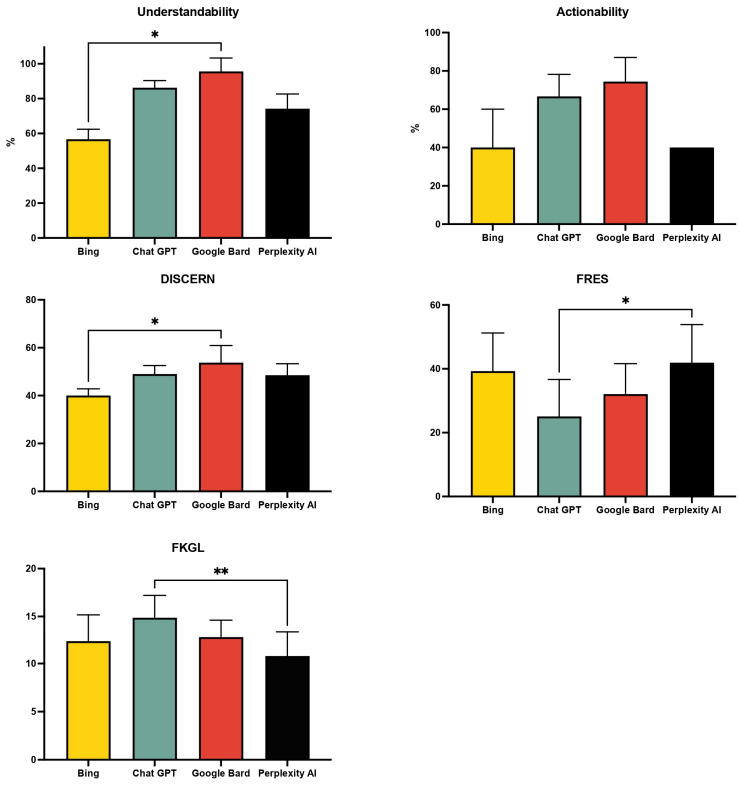

Figure 3. Kruskal–Wallis analysis. Bar graphs represent the median (IQR). Alpha set to 0.05. * $p < 0.05$, and ** $p < 0.01$. (FRES, Flesch–Kincaid Reading Ease Score; and FKGL, Flesch–Kincaid Grade Level).

4. Discussion

In our evaluation of various chat engines, ChatGPT and Google Bard emerged as top performers based on key metrics such as understandability, actionability, and overall quality of responses based on DISCERN and PEMAT scoring. However, it is worth noting that both ChatGPT and Google Bard scored lower in terms of readability as per the FRES and

FKGL assessments. Google Bard's infographics were not recognized within the responses, potentially affecting its FRES and FKGL scores negatively. Despite Perplexity scoring higher in readability, the generated text still maintained an eleventh-grade complexity.

Despite not ranking the highest overall, Perplexity stood out for its extensive use of citations across a wide range of topics. However, it tended to offer repetitive answers despite the diversity of questions posed to it. On the other hand, Google Bard distinguished itself by providing visually engaging and informative graphics to complement its responses, offering a unique form of pictorial education. Additionally, Google Bard demonstrated a high level of detail in its answers. Conversely, Bing, while being the most concise in its responses, scored the lowest overall. Its tendency to offer repetitive answers may have contributed to its lower DISCERN and PEMAT scores, although it did fare better in terms of readability scores.

With the rapid rise of ChatGPT in 2022 and the emergence of various chatbots, their use has become increasingly prevalent in healthcare, with a notable application being their role in answering medical queries [16]. Numerous studies have been carried out to assess the quality, accuracy, and readability of the information generated by different AI chatbots. Musheyev et al. evaluated responses regarding urological malignancies and found high-quality information across different platforms, with Perplexity and Bing scoring the highest and ChatGPT the lowest in terms of the DISCERN score [17]. ChatGPT provided 97% accurate information on cancer when compared against National Cancer Institute's answers [18]. Another study by Seth et a. [19] demonstrated that Bard outperformed ChatGPT and Bing in terms of the DISCERN score when providing information about rhinoplasty.

When analyzing the readability of the responses generated by the chatbots, their scores consistently ranged between eleventh- and college-grade levels [17,20]. This level of readability is significantly higher than the sixth-grade level recommended by the NIH [21]. The discrepancy highlights a barrier for the general public, particularly those with limited health literacy, raising concerns about the understandability and accessibility of AI chatbots for a broader population.

4.1. Clinical Implications

AI chatbots provide instant access to information and can offer support to patients across various settings, from addressing questions about their treatment plans to providing crucial information before a surgery or assistance with the interpretation of lab results. Multiple studies have demonstrated variability in quality scores across different AI chatbots, highlighting the differences in the information they provide across various healthcare topics. Relying on a single chatbot can potentially provide incomplete or inaccurate information. Therefore, patients must be encouraged to validate chatbot-generated advice by consulting healthcare professionals. Furthermore, to improve readability, making use of certain prompts like "Explain it to me in simple terms" may be useful [22].

From a physician's perspective, it is essential to recognize the limitations of these chatbots and engage in thorough discussions with patients to dispel common myths and misconceptions.

4.2. Future Development

As technology advances, the quality and readability of information from chatbots are likely to improve. Out of the three chatbots evaluated, only ChatGPT lacked citations in its responses, highlighting a critical area for improvement. Although many citations were from Mayo Clinic, Cleveland Clinic, Harvard Health Publishing, and WebMD, further progress could be achieved by exclusively referencing from trusted and regulated resources

like those mentioned above. Additionally, providing a user feedback portal to report misinformation could serve as an additional means to maintain accuracy and reliability in the information generated by chatbots.

4.3. Limitations

Our study is limited by the inherent stochastic nature of AI chatbots, as the responses can vary with similar prompts. This reflects their training process, which is based on probabilistic algorithms to generate responses, leading to inconsistencies. Further research is required to assess whether multiple answers to the same question show variability in the scores. Furthermore, the questions were asked sequentially, where the first response could influence subsequent responses. Our goal was to replicate a patient's experience, where multiple questions might be asked during a single interaction.

5. Conclusions

While each chatbot exhibited strengths and weaknesses, ChatGPT and Google Bard excelled in providing comprehensive and actionable responses, with Google Bard offering additional value through visual aids. The answers remained difficult to read across all chatbots relative to the desired reading level. While most of the responses were accurate, some were medically incorrect. The chatbots may be further improved to incorporate references exclusively from trusted sources when it comes to healthcare information, alongside implementing mechanisms to record inaccuracies. Until then, patients and providers should be aware of the various strengths and shortcomings of various chatbots.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/siuj6010013/s1.

Author Contributions: Conceptualization, P.S.; methodology, H.P., A.L. and P.S.; validation, N.N. and P.S.; formal analysis, A.L.; data curation, H.P.; writing—original draft preparation, H.P. and A.L.; writing—review and editing, N.N. and P.S.; visualization, A.L. and H.P.; and supervision, N.N. and P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments: The abstract for this paper was presented at the National Reproductive Health Conference, Philadelphia, USA, on 11 September 2024 and the North Central Section of the AUA on 19 September 2024. The abstract was also presented at the Société Internationale d'Urologie, in New Delhi, on 26 October 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hesse, B.W.; Nelson, D.E.; Kreps, G.L.; Croyle, R.T.; Arora, N.K.; Rimer, B.K.; Viswanath, K. Trust and sources of health information: The impact of the Internet and its implications for health care providers: Findings from the first Health Information National Trends Survey. *Arch. Intern. Med.* **2005**, *165*, 2618–2624. [CrossRef] [PubMed]
2. Calixte, R.; Rivera, A.; Oridota, O.; Beauchamp, W.; Camacho-Rivera, M. Social and demographic patterns of health-related Internet use among adults in the United States: A secondary data analysis of the health information national trends survey. *Int. J. Environ. Res. Public Health* **2020**, *17*, 6856. [CrossRef] [PubMed]
3. Fahy, E.; Hardikar, R.; Fox, A.; Mackay, S. Quality of patient health information on the Internet: Reviewing a complex and evolving landscape. *Australas. Med. J.* **2014**, *7*, 24–28. [CrossRef] [PubMed]

4. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **2020**, *30*, 681–694. [CrossRef]
5. Caldarini, G.; Jaf, S.; McGarry, K. A literature survey of recent advances in chatbots. *Information* **2022**, *13*, 41. [CrossRef]
6. Khanna, A.; Pandey, B.; Vashishta, K.; Kalia, K.; Pradeepkumar, B.; Das, T. A study of today's AI through chatbots and rediscovery of machine intelligence. *Int. J. U-E-Serv. Sci. Technol.* **2015**, *8*, 277–284.
7. Adamopoulou, E.; Moussiades, L. Chatbots: History, technology, and applications. *Mach. Learn. Appl.* **2020**, *2*, 100006. [CrossRef]
8. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
9. Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from language models. *arXiv* **2021**, arXiv:2112.04359.
10. Alkaissi, H.; McFarlane, S.I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **2023**, *15*, e35179. [CrossRef]
11. Gilbert, K.; Cimmino, C.B.; Beebe, L.C.; Mehta, A. Gaps in patient knowledge about risks and benefits of testosterone replacement therapy. *Urology* **2017**, *103*, 27–33. [CrossRef] [PubMed]
12. Discern Online—The Discern Instrument. Available online: http://www.discern.org.uk (accessed on 25 January 2024).
13. Agency for Healthcare Research and Quality. The Patient Education Materials Assessment Tool (PEMAT) and User's Guide. October 2013. Updated November 2020. Available online: https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html (accessed on 25 January 2024).
14. Weil, A.G.; Bojanowski, M.W.; Jamart, J.; Gustin, T.; Lévêque, M. Evaluation of the quality of information on the Internet available to patients undergoing cervical spine surgery. *World Neurosurg.* **2014**, *82*, e31–e39. [CrossRef] [PubMed]
15. Readability Formulas—Readability Scoring System. Available online: https://readabilityformulas.com/readability-scoring-system.php#formulaResults (accessed on 29 January 2024).
16. Liu, J.; Wang, C.; Liu, S. Utility of ChatGPT in clinical practice. *J. Med. Internet Res.* **2023**, *25*, e48568. [CrossRef] [PubMed]
17. Musheyev, D.; Pan, A.; Loeb, S.; Kabarriti, A.E. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur. Urol.* **2024**, *85*, 13–16. [CrossRef] [PubMed]
18. Johnson, S.B.; King, A.J.; Warner, E.L.; Aneja, S.; Kann, B.H.; Bylund, C.L. Using ChatGPT to evaluate cancer myths and misconceptions: Artificial intelligence and cancer information. *JNCI Cancer Spectr.* **2023**, *7*, pkad015. [CrossRef]
19. Seth, I.; Lim, B.; Xie, Y.; Cevik, J.; Rozen, W.M.; Ross, R.J.; Lee, M. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: An observational study. *Aesthetic Surg. J. Open Forum* **2023**, *5*, ojad084. [CrossRef] [PubMed]
20. Pan, A.; Musheyev, D.; Bockelman, D.; Loeb, S.; Kabarriti, A.E. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. *JAMA Oncol.* **2023**, *9*, 1437–1440. [CrossRef] [PubMed]
21. Weiss, B.D. Health literacy. *Am. Med. Assoc.* **2003**, *253*, 358.
22. Pividori, M. Chatbots in Science: What Can ChatGPT Do for You? Available online: https://www.nature.com/articles/d41586-024-02630-z (accessed on 4 December 2024).