Responses of AI Chatbots to Testosterone Replacement Therapy: Patient's Beware!

Herleen Pabla , Alyssa Lange , Nagalakshmi Nadiminty , Puneet Sindhwani *

Posted Date: 2 October 2024

doi: 10.20944/preprints202410.0152.v1

Keywords: artificial intelligence; misinformation; testosterone

 doi:10.20944/preprints202410.0152.v1

Article

Responses of AI Chatbots to Testosterone Replacement Therapy: Patient's 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, United States
[2] University of Toledo Medical School, University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, United States
* Correspondence: Department of Urology, The University of Toledo College of Medicine and Life Sciences, Toledo, OH 43614, United States. puneet.sindhwani@utoledo.edu, Tel: +14193833578

Abstract: Purpose: Using Chatbots to seek healthcare information is becoming more popular. Misinformation and gaps in knowledge exist regarding the risks and benefits of Testosterone Replacement Therapy (TRT). We aimed to assess and compare the quality and readability of responses generated by four AI chatbots. **Materials and Methods:** ChatGPT, Google Bard, Bing Chat, and Perplexity AI were asked the same eleven questions regarding Testosterone Replacement Therapy. 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 Flesch-Kincaid Grade Level (FKGL). Kruskal-Wallis statistics were completed using GraphPad Prism V10.1.0. **Results:** Google Bard received the highest DISCERN and PEMAT. Perplexity AI received the highest FRES and best FKGL. 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. **Conclusion:** ChatGPT and Google Bard were 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.

Keywords: artificial intelligence; misinformation; testosterone

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]. It utilizes 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].

 doi:10.20944/preprints202410.0152.v1

Testosterone Replacement Therapy (TRT) is a subject clouded by misinformation and there are existing gaps in knowledge about the benefits and risks among patients [10]. With the advent of multiple AI chatbots and their increasing use, it becomes imperative to further analyze the quality, accuracy and readability of 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 the AI chatbot responses to Testosterone Replacement Therapy. Therefore, the objective of our study is to assess and compare the accuracy, quality and readability of responses generated by four popular AI chatbots - Google Bard, Chat GPT3.5, Perplexity AI, and Bing Chat concerning TRT.

Materials and Methods

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 to simulate a patient's perspective regarding TRT as seen in Figure 1. The questions were asked in sequential order and remained the same across all chatbot engines. The list of all questions and answers generated by all chatbots is provided in Supplementary Data 1.

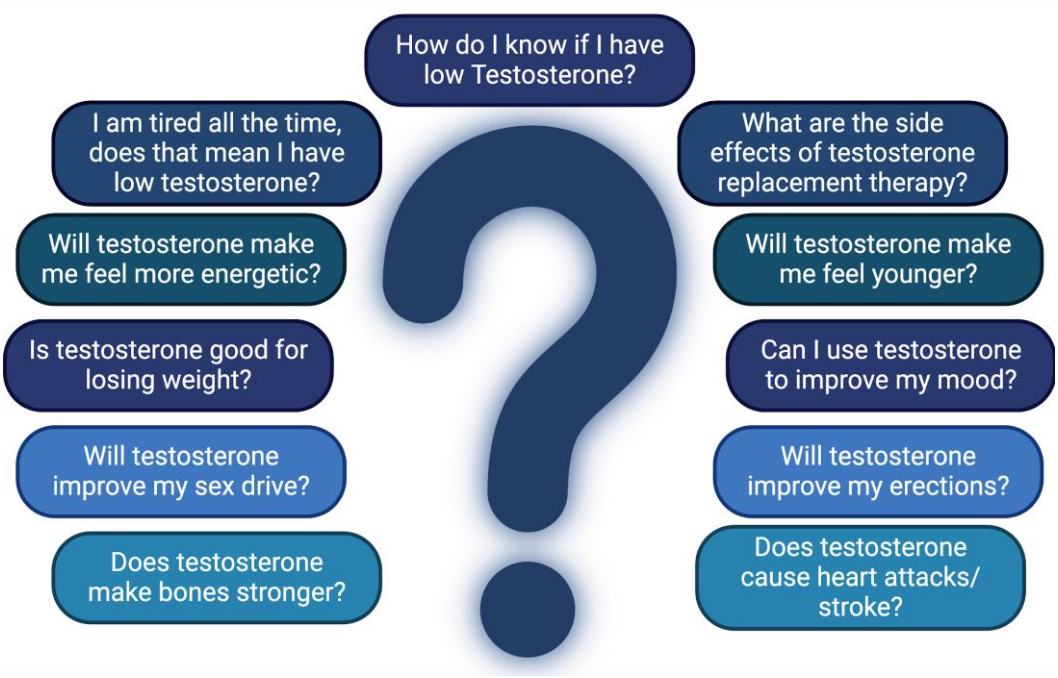

Figure 1. The questions posed to each chatbot stimulating a patient's perspective.

To analyze for accuracy, the responses were compared to published and peer-reviewed guidelines from the AUA. The responses were evaluated by four independent reviewers - 1 board-certified Urologist, 1 PhD Urology, MS III and IV. The primary outcomes were the quality of consumer health information based on the validated DISCERN instrument [11] and the understandability and actionability via the Patient Education Materials Assessment Tool (PEMAT) [12]. 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 - 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" [13]. PEMAT scores range between 0%-100%, with higher scores indicating a higher level of understandability and actionability [12].

Readability was assessed using the Flesch-Kincaid Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL) using an online readability calculator [14]. FRES ranged from 0 to 100 with a higher score corresponding to a text that is easier to read (recommended score ≥ 70) and FKGL measured the grade level necessary to comprehend the text (recommended score of ≤ 8). Kruskal-

 doi:10.20944/preprints202410.0152.v1

Wallis statistics were completed using GraphPad Prism V10.1.0. An example question and answer generated by ChatGPT is given. See Figure 2.

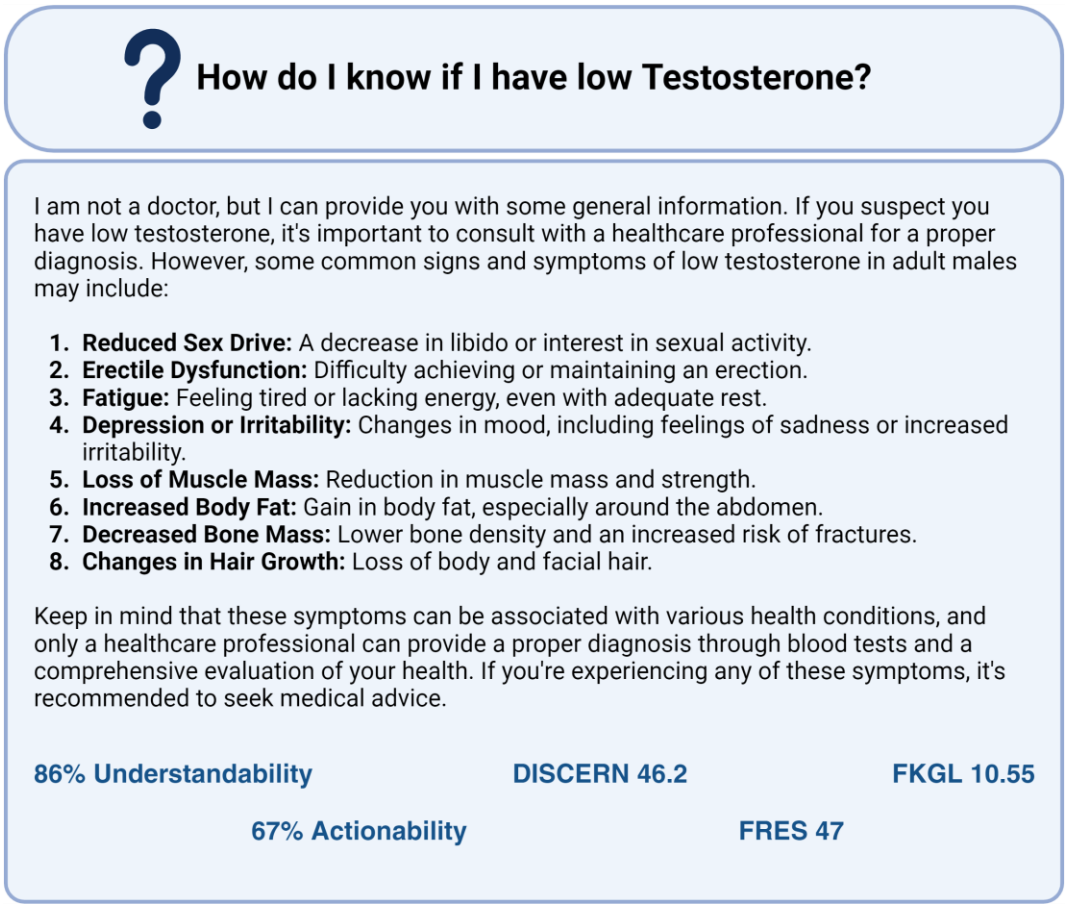

Figure 2. ChatGPT's response to one of the questions and summary of the scores.

Results

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

Quality of information

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

Understandability and actionability of information

For PEMAT understandability scores, Bing AI had the lowest 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 had Bing Chat and Perplexity at the lowest with 40% (20-60%) and (40%-40%) respectively. ChatGPT received a score of 67% (60-80%) and Google Bard again had the highest score of 74% (60-83.3%).

Readability

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

 doi:10.20944/preprints202410.0152.v1

Perplexity AI had the best FKGL score of 10.8, followed by Bing Chat at 12.3, Google Bard received 12.7, and the worst score was for Chat GPT at 14.9. Table 1 shows a summary of the scores received by each of the four AI chatbots.

Table 1. Summary of scores of each of the four AI chatbots.

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

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

 doi:10.20944/preprints202410.0152.v1

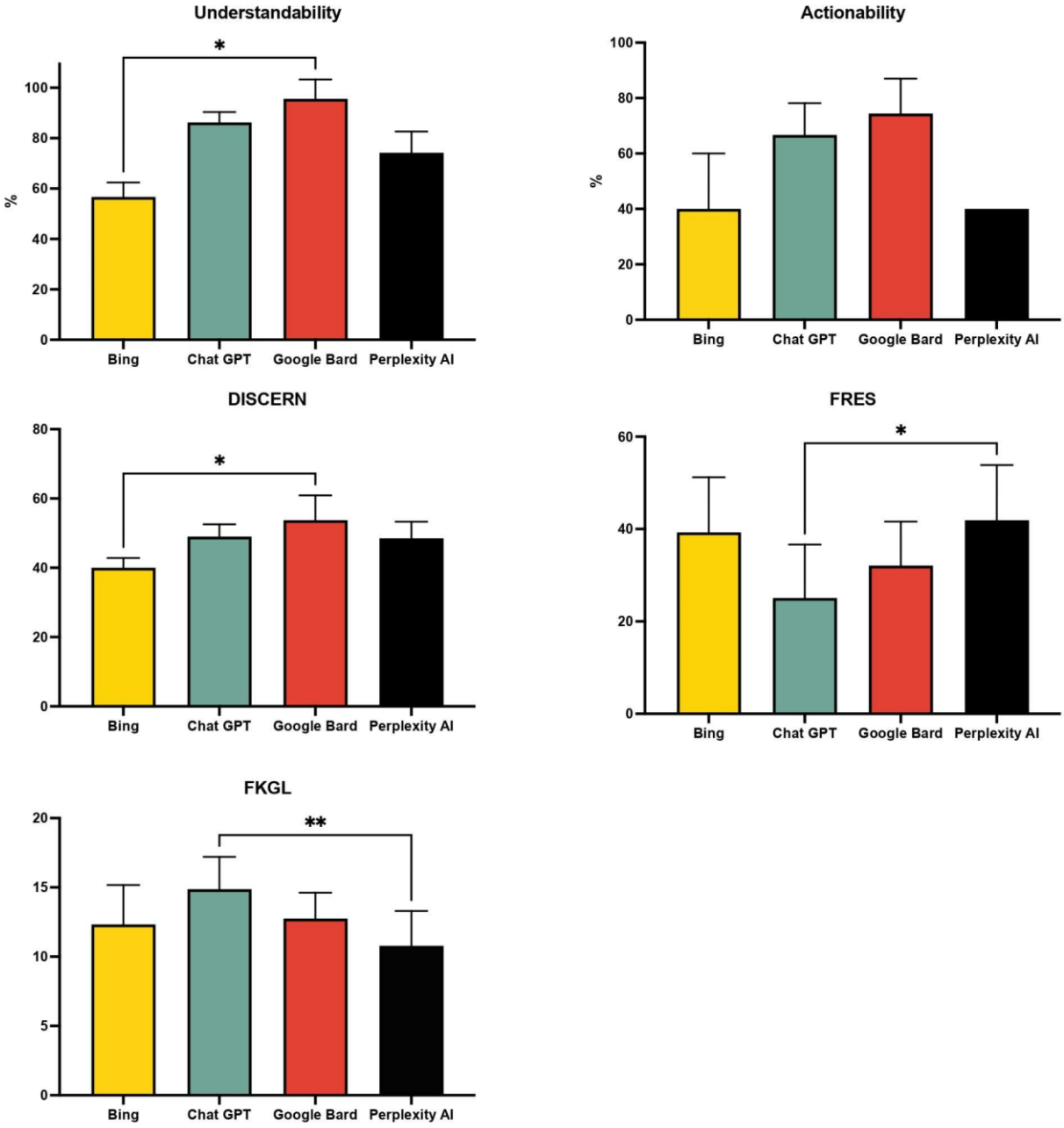

Figure 3. Kruskal-Wallis analysis. Bar graphs represent mean with standard deviation. Alpha set 0.05.

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's worth noting that both ChatGPT and Google Bard scored lower in terms of readability and grade level as per the FRES and FKGL assessments. Google Bard's infographics may not be recognized within 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

 doi:10.20944/preprints202410.0152.v1

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 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 and the FKGL score.

Overall, while each chat engine exhibited strengths and weaknesses, ChatGPT and Google Bard excelled in providing comprehensive and actionable responses, with Google Bard offering additional value through visual aids, despite challenges in recognizing certain types of content.

Conclusions

These findings highlight the need for patients and providers to know the various shortcomings and strengths of various commonly used AI Chatbots. While most of the responses were accurate some responses were medically incorrect. As the technology improves, the algorithms might increase in accuracy. Till then patients should beware of the wrong information that might be generated by the AI Chatbots.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, P.S.; methodology, H.P., A.L., P.S.; validation, N.N., P.S.; formal analysis, A.L.; data curation, H.P.; writing—original draft preparation, H.P., A.L.; writing—review and editing, N.N., P.S.; visualization, A.L., H.P.; supervision, N.N., 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.

Conflict of interest: The authors declare that they have nothing to disclose.

References

1. Hesse, B. W., Nelson, D. E., Kreps, G. L., et al. 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. *Archives of internal medicine*. 2005; 165(22):2618-2624.
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. *International Journal of Environmental Research and Public Health*. 2020; 17(18):6856.
3. Fahy, E., Hardikar, R., Fox, A., & Mackay, S. Quality of patient health information on the Internet: reviewing a complex and evolving landscape. *The Australasian medical journal*. 2014; 7(1):24–28.
4. Floridi, L., & Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*. 2020; 30:681-694.
5. Caldarini, G., Jaf, S., & McGarry, K. A literature survey of recent advances in chatbots. *Information*. 2022; 13(1):41.
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. *International Journal of u-and e-Service, Science and Technology*. 2015; 8(7):277-284.
7. Adamopoulou, E., & Moussiades, L. Chatbots: History, technology, and applications. *Machine Learning with Applications*. 2020; 2:100006.
8. Ji, Z., Lee, N., Frieske, R., et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023; 55(12):1-38.
9. Weidinger, L., Mellor, J., Rauh, M., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*. 2021.
10. 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.
11. Discern online - THE DISCERN Instrument. Accessed January 25, 2024. http://www.discern.org.uk/discern_instrument.php

 doi:10.20944/preprints202410.0152.v1

12. Agency for Healthcare Research and Quality. The Patient Education Materials Assessment Tool (PEMAT) and user's guide. October 2013. Updated November 2020. Accessed January 25, 2024. https://www.ahrq.gov/health-literacy/patient-education/pemat-p.html
13. Weil AG, Bojanowski MW, 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(1-2):e31-e39.
14. Readability formulas - READABILITY SCORING SYSTEM. Accessed January 29, 2024. https://readabilityformulas.com/readability-scoring-system.php#formulaResults