Cureus
Part of SPRINGER NATURE

Open Access Original Article

DOI: 10.7759/cureus.

Unlocking Health Literacy: The Ultimate Guide to Hypertension Education from ChatGPT vs. Google Gemini

Thomas J. Lee [1], Daniel J. Campbell [2], Shriya Patel [1], Afif Hossain [3], Navid Radfar [1], Emaad Siddiqui [3], Julius M. Gardin [3]

1. Department of Medicine, Rutgers University New Jersey Medical School, Newark, USA 2. Otolaryngology–Head and Neck Surgery, Thomas Jefferson University Hospital, Philadelphia, USA 3. Department of Cardiology, Rutgers University New Jersey Medical School, Newark, USA

Corresponding author: Thomas J. Lee, tl467@njms.rutgers.edu

Abstract

Background

Google Gemini represents the latest advances in the realm of artificial intelligence (AI) and has garnered attention due to its capabilities similar to the increasingly popular ChatGPT. Accurate dissemination of information on common conditions such as hypertension is critical for patient comprehension and management. Despite the ubiquity of AI, comparisons between ChatGPT and Gemini remain largely unexplored.

Methods

ChatGPT and Gemini were asked 52 questions derived from the American College of Cardiology's (ACC) frequently asked questions on hypertension, each time following a specified prompt. Prompts included: no prompting (Form 1), patient-friendly prompting (Form 2), physician-level prompting (Form 3), and prompting for statistics/references (Form 4). Responses were scored as incorrect, partially correct, or correct. Flesch-Kincaid (FK) grade level and word count were recorded.

Results

Across all forms, scoring frequencies were: 23 (5.5%) incorrect, 162 (38.9%) partially correct, and 231 (55.5%) correct. ChatGPT showed higher rates of partially correct answers than Gemini ($p=0.0346$). Physician-level prompts resulted in higher word count across both platforms ($p<0.001$). ChatGPT showed a higher FK grade level ($p=0.033$) in physician-friendly prompting. Gemini exhibited a significantly higher mean word count ($p<0.001$); however ChatGPT had a higher FK grade level across all forms ($p>0.001$).

Conclusion

To our knowledge, this study is the first to compare cardiology-related responses from ChatGPT and Gemini, two of the most popular AI chatbots. The grade level for most responses were collegiate-level, which was above average for the National Institutes of Health (NIH) recommendations, but on-par with most online medical information. Both chatbots responded with a high degree of accuracy, with inaccuracies being rare. Therefore, it is reasonable that cardiologists suggest either chatbot as a source of supplementary education.

Categories: Cardiology, Medical Education, Healthcare Technology
Keywords: google gemini, chatgpt, hypertension, patient education, artificial intelligence

Introduction

Hypertension is the most common cardiovascular disease and leading cause of cardiovascular mortality worldwide [1,2]. Over half of the world's population has been diagnosed with hypertension, with a much greater percentage suspected, but not yet diagnosed with hypertension [3]. Furthermore, hypertension is a systemic disease contributing to adverse effects in multiple organ systems and in overall lifestyle. Patient education plays an especially pivotal role in managing hypertension, given that disease treatment is inherently multi-factorial, involving medications, diet, exercise, life stressors, etc [4].

According to the National Cancer Institute's (NCI) Health Information National Trends Survey (HINTS), 84.6% of the US adult population used the Internet to look for health or medical information in 2022, with a number expected to rise in the coming decade [5]. The literature on the prevalence of utilizing artificial intelligence (AI) chatbots for medical education is poorly defined; however, the trend of utilization of AI chatbots has been steadily rising even after its initial exponential rise.

ChatGPT, an AI chatbot developed by OpenAI in November of 2022, has quickly gained widespread attention. From its inception, the site took 5 days to reach 1 million users and within two months it had surpassed 100 million users [6]. In response to the rise of AI chatbots, Google released Gemini, a large language model similar to that of ChatGPT, on May 10, 2023. Within a few months, many speculated Gemini would be the primary competitor to ChatGPT [7,8].

Given AI's burgeoning popularity and its potential for disseminating health information, evaluating the quality and accuracy of ChatGPT and Gemini is of paramount importance. We aimed to critically assess ChatGPT's responses to queries about one of the world's most common diseases, hypertension. This study focuses on the accuracy, comprehensibility, and appropriateness of using AI responses for patient education. The good of the study is to guide cardiologists and healthcare professionals in understanding the benefits and potential limitations of AI for patient education.

Materials And Methods

OpenAI's ChatGPT and Google's Gemini chatbots were prompted four times, then asked 52 questions from the 2017 American College of Cardiology's (ACC) frequently asked questions on hypertension [9]. ChatGPT version 3.5 and Gemini version 1.0 (formerly known as Google Bard) were used for all responses. All questions were asked between the dates of September 6, 2023 and September 7, 2023.

Prompts were as follows: no prompt (Form 1), patient-friendly prompt (Form 2), physician-level prompt (Form 3), and prompting for statistics/references (Form 4). The prompts used are in Table 1. Responses were reviewed and scored as incorrect, partially correct, correct, or correct with references (perfect). Incorrect responses were designated if the response included any incorrect information or if responses included less than 50% of information from the ACC response answers. Partially correct answers included responses that had no incorrect information and included 50% - 99% of the information from the ACC responses. Correct responses included all information from the ACC responses with any extra information being correct. Perfect responses included responses that met criteria for correct responses and included references and/or statistics in the response. Proportions of responses at differing scores were compared using chi-square analysis. Tests were performed with an alpha set at 0.05.

For each response, the number of words, sentences, and syllables were collected to compute a Flesh-Kincaid (FK) Grade level. This metric estimates the United States educational grade level required to understand the response, with higher grade levels indicating more complex language usage and is defined as:

$$0.39\left(\frac{words}{sentences}\right)+11.8\left(\frac{syllables}{words}\right)-15.59$$

Values vary from 0-20, with the numerical value corresponding with the reading grade level (e.g., 12 would equal grade level 12). Significance between forms was calculated using a one-way ANOVA with an alpha of 0.05. Additionally, response length was recorded and significance was analyzed with a one-way ANOVA and an alpha set at 0.05. Significance for statistical analyses was set at $p < 0.05$. Statistics were run using Prism 10.0.2.

Results

Across all forms, scoring frequencies for ChatGPT were: 9 (4.33%) incorrect, 92 (44.23%) partially correct, and 107 (51.44%) correct. Scoring frequencies for Gemini were: 14 (6.73%) incorrect, 70 (33.65%) partially correct, and 124 (59.62%) correct. Chi-squared analysis revealed the proportions of responses categorized as correct did not significantly differ between ChatGPT vs Gemini (p=0.11). However, ChatGPT was more likely to give a partially correct response when compared to Gemini (p=0.035) (Figure *1*).

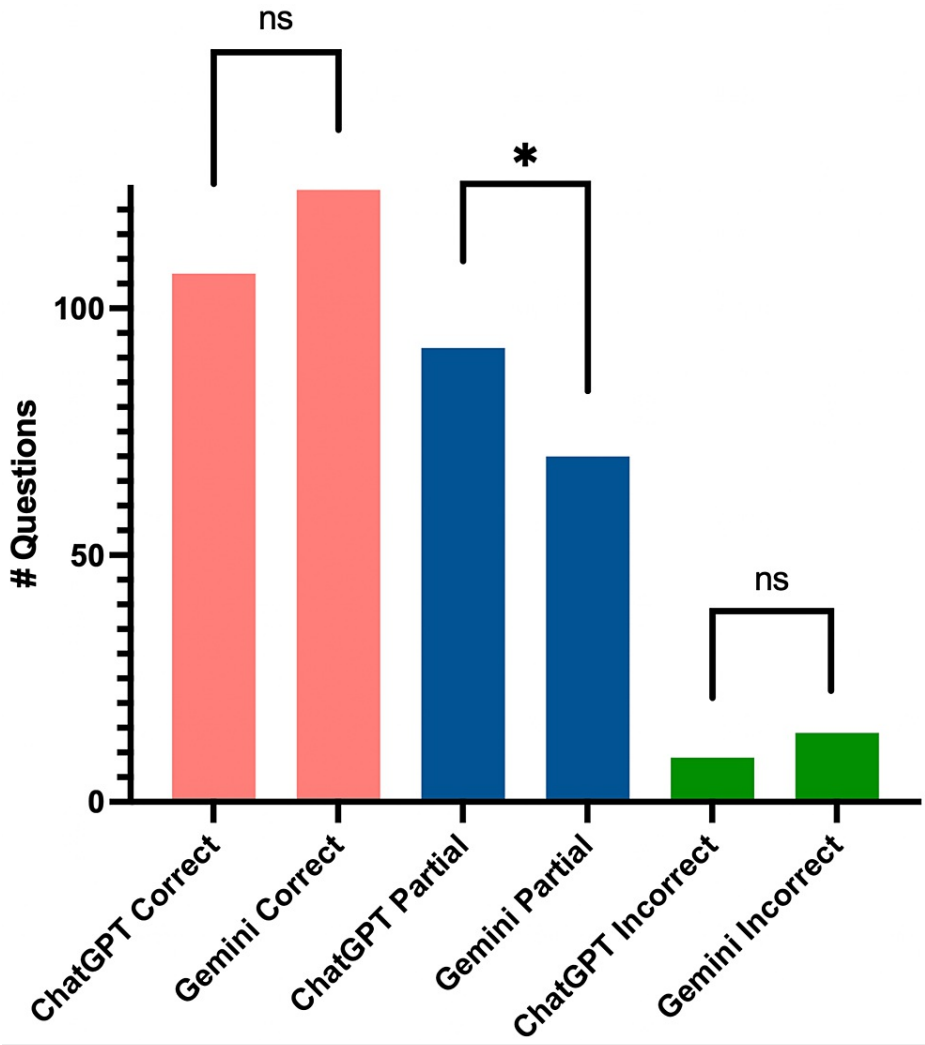

FIGURE 1: Correct, Partially Correct, and Incorrect Answers in ChatGPT and Gemini Responses

Each bar shows the total number of correct, partially correct, or incorrect answers between all forms.
Abbreviations: ns = no significance. * = p<0.05.

FK scores for ChatGPT and Gemini can be found in Figure *2*. ChatGPT's mean FK grade reading level was as follows: Form 1 at 16.20 (± 4.36), Form 2 at 15.15 (± 4.25), Form 3 at 17.37 (± 4.20), and Form 4 at 14.94 (± 5.17). In ChatGPT responses, a significant difference was found between Form 3 and 4's grade reading level (p=0.033). Gemini's mean FK grade reading level was as follows: Form 1 at 13.69 (± 3.92), Form 2 at 13.38 (± 3.27), Form 3 at 13.50 (± 3.67), and Form 4 at 13.44 (± 3.37). There was no significant difference in reading level between forms for Gemini's responses. Overall ChatGPT's responses had a higher grade reading level (15.92 ± 4.58) than Gemini's responses (13.50 ± 3.54) (p<0.0001).

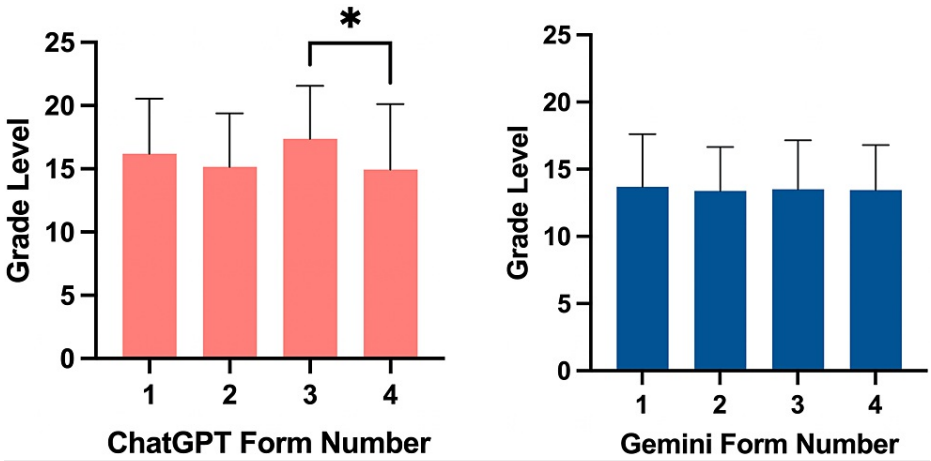

FIGURE 2: Grade Reading Level for ChatGPT and Gemini Responses

* = p<0.05.

Word count for ChatGPT and Gemini can be found in Figure *3*. ChatGPT's mean word count was as follows: Form 1 at 19.00 (± 6.74), Form 2 at 19.00 (± 8.84), Form 3 at 18.00 (± 6.30), and Form 4 at 20.50 (± 14.07). Gemini's mean word count was as follows: Form 1 at 27.50 (± 28.02), Form 2 at 38.00 (± 22.81), Form 3 at 27.50 (± 19.41), and Form 4 at 56.00 (± 40.45). Overall Gemini's responses had a higher word (44.43 ± 30.82) than ChatGPT's responses (21.08 ± 9.75) (p<0.0001).

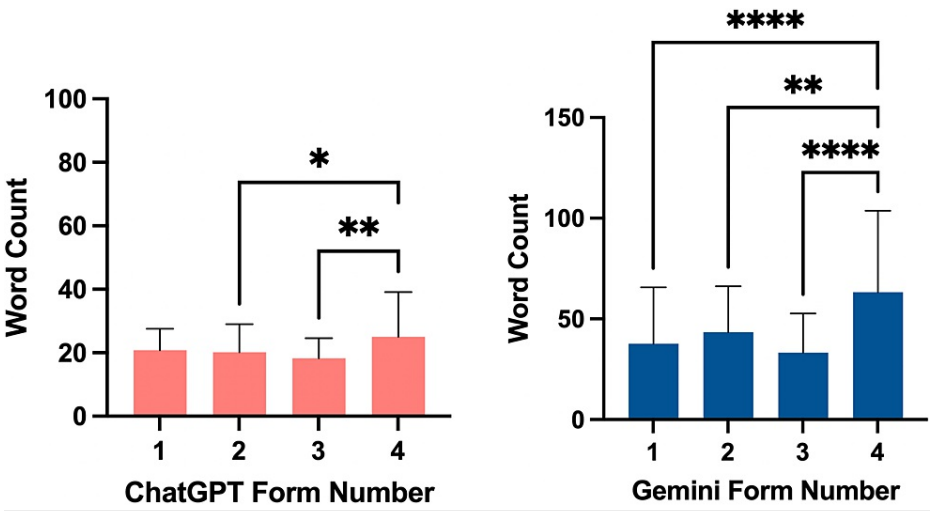

FIGURE 3: Word Count for ChatGPT and Gemini Responses

* = p<0.05; ** = p<0.01; *** = p<0.0001.

Discussion

From the inception of Google Gemini, comparisons to ChatGPT were made, and there have been many speculation about which AI chatbot would be more accurate [10,11]. To date, few studies have objectively compared the accuracy of responses, with even fewer studies focusing on the medical field [12]. To our knowledge, this is the first study to compare the performance of two of the most popular AI chatbots, ChatGPT and Gemini, on cardiology related topics.

Overall, both ChatGPT and Gemini provided accurate, but often partially complete, responses when responding to ACC's frequently asked questions about hypertension. Even though only half the answers were deemed entirely "correct," this result was still seen positively. The AI chatbots' replies often contained more than 50% of the information, typically lacking just one element from the ACC's answers. The responses that would signify a large deficiency - incorrect or incomplete (greater than 50% missing) information - were only

present in 5.5% of responses. This result was on par with many other studies that examined artificial intelligent chatbot responses, generally ranging from 1-5% incorrect responses [13-16]. ChatGPT gave more partially correct answers than Gemini, while Gemini exhibited a non-significant trend to provide more correct responses than did ChatGPT. This could be in part because ChatGPT's mean responses were 23 words shorter than Gemini's responses, thus leaving less room for information.

ChatGPT had a higher mean grade reading level than Gemini, with an FK score of 15.92 versus 13.50, respectively. Although ChatGPT's answers were less accurate, they were more succinct and used a higher grade reading level. The National Institutes of Health (NIH) recommends patient education material should be written at an 8th grade reading level, which is lower than both ChatGPT's approximate grade level (grade 15 - collegiate level), and Gemini's approximate grade reading level (grade 13 - collegiate level) [17]. However, ChatGPT and Gemini's grade levels are quite similar to many online sources of cardiology material. Academic websites pertaining to atrial fibrillation had a mean grade level of 13.05, while non-academic sites had a mean average of 11.64 [18]. This finding was mirrored in other medical specialties' online reading material [19-22]. Therefore, while the two chatbots responded above the NIH recommended grade level, the responses were on-par with most online resources.

ChatGPT consistently had a lower average word count in its responses compared to Gemini, as noted earlier. Similar trends have been observed in other studies comparing the two chatbots' performance on health literacy, hinting that Gemini may naturally provide lengthier responses [23]. Notably, the word count for both chatbots remained fairly consistent across various query types with the exception of Form 4, which involves requesting statistics or research. This variation is likely due to the nature of the prompt, as requesting data and references typically necessitates the inclusion of more detailed information, such as citations, statistical figures, or mathematical equations.

While this study assesses responses objectively, it has its limitations, including the assumption of accurate patient inquiries. We did not assess the chatbots' reactions to false information. Also, patients have myriad ways to ask questions, potentially leading to responses not reviewed in this study. Future research should broaden the scope of inquiries and analyze the chatbots' handling of erroneous inputs.

Conclusions

The analysis shows that AI chatbots like ChatGPT and Gemini can be valuable tools for augmenting patient education on topics such as hypertension. Both have demonstrated a strong ability to provide accurate answers. They might not include every nuance that the ACC offers, but they generally convey the necessary information with few errors. Therefore, it's sensible for medical professionals to suggest using ChatGPT or Gemini as educational resources. Nevertheless, one should recognize the minor possibility of encountering inaccuracies.

Additional Information

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Jaeger BC, Chen L, Foti K, et al.: Hypertension Statistics for US Adults: An Open-Source Web Application for Analysis and Visualization of National Health and Nutrition Examination Survey Data. Hypertens (Dallas, Tex. 19792023, 80:1311-1320. 10.1161/HYPERTENSIONAHA.123.20900
2. Mills KT, Stefanescu A, He J: The global epidemiology of hypertension . Nat Rev Nephrol. 2020, 16:223-237. 10.1038/s41581-019-0244-2
3. NCD Risk Factor Collaboration (NCD-RisC: Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants. Lancet (London, England. 2021, 398:957-980. 10.1016/S0140-6736(21)01330-1
4. Waeber B, Brunner HR: The multifactorial nature of hypertension: the greatest challenge for its treatment? . Journal of Hypertension. 2001, 19:9-16.

5. Health Information National Trends Survey. (2022). Accessed: December 7, 2023: https://hints.cancer.gov/view-questions/question-detail.aspx?PK_Cycle=14&qid=1926.
6. ChatGPT reaches 100 million users two months after launch February 2, 2023. Accessed May 2 . (2023). Accessed: Accessed September 19, 2023: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users- open-ai-fastest-growing-app.
7. AI bros are at war over declarations that Google's upcoming Gemini AI model smashes OpenAI's GPT-4 . (2024). Accessed: March 25, 2024: https://www.businessinsider.com/google-gemini-ai-model-smashes-gpt4-says-semianalysis-2023-8.
8. OpenAI Rages at Report that Google's New AI Crushes GPT-4 . (2024). Accessed: March 25, 2024: https://futurism.com/the-byte/openai-report-google-ai-gpt-4.
9. American College of Cardiology Frequently Asked Question on Hypertension . (2017). Accessed: Accessed January 22, 2024: https://www.acc.org/guidelines/hubs/high-blood-pressure/frequently-asked-questions.
10. Google Gemini vs. ChatGPT: Is Gemini Better Than ChatGPT? . (2024). Accessed: Februry 8, 2024: https://www.techrepublic.com/article/google-gemini-vs-chatgpt/.
11. AI Showdown: ChatGPT Vs. Google's Gemini - Which Reigns Supreme? February 26 . (2024). Accessed: March 29, 2024: https://bernardmarr.com/ai-showdown-chatgpt-vs-googles-gemini-which-reigns-supreme/.
12. Carlà MM, Gambini G, Baldascino A, et al.: Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. March. 2024, 10.1136/bjo-2023-325143
13. Soto-Chávez MJ, Bustos MM, Fernández-Ávila DG, Muñoz OM: Evaluation of information provided to patients by ChatGPT about chronic diseases in Spanish language. Digit Heal. 2024, 10:20552076231224604. 10.1177/20552076231224603
14. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S: Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer?. Front Oncol. 2023, 13:1256459. 10.3389/fonc.2023.1256459
15. Campbell DJ, Estephan LE, Mastrolonardo E V, Amin DR, Huntley CT, Boon MS: Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin sleep Med JCSM Off Publ Am Acad Sleep Med. 2023, 19:1989-1995. 10.5664/jcsm.10728
16. Campbell DJ, Estephan LE, Sina EM, et al.: Evaluating ChatGPT Responses on Thyroid Nodules for Patient Education. Thyroid. 2024, 34:371-377. 10.1089/thy.2023.0491
17. Rooney MK, Santiago G, Perni S, et al.: Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. J patient Exp. 2021, 8:2374373521998847. 10.1177/2374373521998847
18. Siddiqui E, Shah AM, Sambol J, Waller AH: Readability Assessment of Online Patient Education Materials on Atrial Fibrillation. Cureus. 2020, 12:10397. 10.7759/cureus.10397
19. Huang G, Fang CH, Agarwal N, Bhagat N, Eloy JA, Langer PD: Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmol. 2015, 133:449-454. 10.1001/jamaophthalmol.2014.6104
20. Gupta R, Adeeb N, Griessenauer CJ, et al.: Evaluating the complexity of online patient education materials about brain aneurysms published by major academic institutions. J Neurosurg. 2017, 127:278-283. 10.3171/2016.5.JNS16793
21. Svider PF, Agarwal N, Choudhry OJ, et al.: Readability assessment of online patient education materials from academic otolaryngology-head and neck surgery departments. Am J Otolaryngol. 2013, 34:31-35. 10.1016/j.amjoto.2012.08.001
22. Singh SP, Qureshi FM, Borthwick KG, Singh S, Menon S, Barthel B: Comprehension Profile of Patient Education Materials in Endocrine Care. Kansas J Med. 2022, 15:247-252. 10.17161/kjm.vol15.16529
23. Amin KS, Mayes LC, Khosla P, Doshi RH: Assessing the Efficacy of Large Language Models in Health Literacy: A Comprehensive Cross-Sectional Study. Yale J Biol Med. 2024, 97:17-27. 10.59249/ZTOZ1966