

Evaluating the proficiency of ChatGPT-3.5, ChatGPT-4o and Google Gemini in responding to frequently asked questions on gender affirming care by transgender people

Selin Tekin¹ , Banu Ertürk¹ , Seda Hanife Oğuz¹ , Bülent Okan Yıldız¹ 

¹Division of Endocrinology and Metabolism, Department of Internal Medicine, Faculty of Medicine, Hacettepe University, Ankara, Türkiye

Abstract

Background: Transgender and gender-diverse individuals encounter stigma, discrimination, and healthcare access barriers, leading them to seek online information, including from large language models like ChatGPT and Google Gemini. This study compares their responses on gender-affirming care.

Methods: This cross-sectional study evaluated responses from ChatGPT-3.5, ChatGPT-4o, and Google Gemini to 18 commonly asked questions on gender-affirming care, derived from current guidelines and clinical practice inquiries. Two endocrinologists independently assessed guideline adherence, while response reliability and quality were evaluated using the modified DISCERN scale and the Global Quality Scale, respectively. Hallucination tendency was rated using a Likert scale, and readability was analyzed using the Flesch Reading Ease, Flesch-Kincaid Grade Level, and Gunning Fog Index.

Results: The highest guideline compatibility was observed with ChatGPT-4o [81.1%], followed by ChatGPT-3.5 [77.2%] and Google Gemini [63.8%]. For quality ChatGPT-4o scored highest [4.8±0.3], ChatGPT-3.5 slightly lower [4.5±0.5], and Gemini significantly lower [2.4±0.5]. ChatGPT-4o and ChatGPT-3.5 showed similar quality [p=0.19] but both outperformed Gemini [p<0.001]. ChatGPT-4o had the greatest tendency to hallucinate. Overall, a high school education was needed to understand the responses. In terms of readability, Gemini's answers were easiest, followed by ChatGPT-3.5 and ChatGPT-4o [p<0.001].

Conclusions: ChatGPT-4o demonstrated the highest accuracy, quality, and guideline adherence in gender-affirming care, outperforming ChatGPT-3.5 and Google Gemini. However, its greater hallucination tendency necessitates caution in medical use. While Google Gemini had better readability, its lower compatibility and quality scores reduce its effectiveness in transgender care. These inaccuracies highlight that AI should not replace expert healthcare guidance.

Keywords: artificial intelligence, ChatGPT, Gemini, gender affirming care, transgender, endocrinology

Introduction

Transgender and gender diverse people encounter significant levels of stigma and discrimination in their daily lives compared to cisgender individuals [1,2].

They may find it challenging to open up to peers or parents, even doctors, or they may have already faced negative reactions when attempting to do so [2]. Transgender individuals face several challenges when seeking gender-affirming healthcare, including

Corresponding author: Melih Oral • Email: melihoral12@gmail.com

Received: October 20, 2025 Accepted: July 16, 2026 Published: September 30, 2026

Copyright © 2026 The Author(s). Published by Hacettepe University Faculty of Medicine. This is an open access article distributed under the [Creative Commons Attribution License \(CC BY\)](#), which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

concerns related to hormone therapy, discrimination, and confidentiality [3]. Furthermore, many medical providers have received limited or no training on how to effectively work with transgender and gender diverse individuals [4], which may lead transgender people to question whether they will receive competent care. Consequently, trans and gender diverse individuals are known to avoid healthcare services due to concerns about encountering difficulties in reaching specialized professionals in transgender health because of the scarcity of healthcare providers experienced in this field [5]. Moreover, discrimination in the society may lead to unemployment or low income, making it harder to access healthcare [2].

Individuals facing limited access to healthcare services often turn to online sources for health information. Transgender and gender diverse individuals, in particular, demonstrate a higher tendency to search for health information online compared to cisgender individuals [6]. Large language models [LLMs] like ChatGPT and Google Gemini are a specialized type of artificial intelligence [AI] algorithm crafted for natural language processing tasks, with a central focus on predicting word sequences by taking into account the surrounding context. ChatGPT is an AI language model developed by OpenAI, based on the generative pre-trained transformer architecture [7]. ChatGPT is pre-trained on an extensive dataset that encompasses data from various online sources, such as websites, articles, books, and other publicly available texts [8]. Google Gemini is a family of large language models [LLMs] developed by Google DeepMind and open-sourced in December 2023. These models are multimodal, capable of understanding and processing information from various formats, including text, code, images, audio, and video [9]. Gemini models can be accessed through a Google account. The accessibility and open-access nature of these LLMs render them an appealing source of information for both patients and healthcare professionals. However, it is crucial not to underestimate current limitations, such as the absence of appropriate references; and to check whether the information is up-to-date, accurate, free of hallucinations, and unbiased [10].

ChatGPT and Google Gemini along with other similarly trained AI systems are subject to potential biases and dissemination of harmful misinformation, including stereotypes, gender norms, harmful dialogue, and deleterious opinions [11]. Transgender individuals may

already be marginalized and stigmatized in various aspects of life and they are susceptible to systemic bias [12,13]. LLMs' comprehension of gender-affirming care and gender identity may be influenced by bias and misinformation, contingent upon factors such as the platform's training data, source code, and reinforced learning experiences. An AI system therefore require examination to ascertain its perspectives and accuracy regarding gender-affirming medical care, in order to evaluate its suitability as an educational resource.

AI systems, despite all these considerations, have nevertheless been widely used to search for answers to numerous medical questions. We considered that transgender individuals could also seek answers to their questions on ChatGPT or Google Gemini due to various reasons such as discrimination, stigma, and difficulties accessing medical services. Thus, we aimed to compare responses of ChatGPT-3.5, ChatGPT-4o and Google Gemini to common questions asked by transgender individuals with existing guideline information.

Methods

Study design

This is a cross-sectional non-human subject study, posing 18 commonly asked questions related to gender affirming care of transgender individuals to ChatGPT-3.5, ChatGPT-4o and Google Gemini. The questions were developed based on current guidelines [14,15], and the inquiries commonly asked by transgender individuals to us as experienced endocrinologists during our endocrinology outpatient clinical practice with transgender individuals.

The questions were categorized into three sections (Table 1);

1. Gender-affirming hormonal therapy [GAHT, n=12]
2. Adverse outcomes and long-term care [n=4]
3. Gender-affirming surgery [n=2]

A prompting question was posed first as 'Please answer the following questions on gender affirming care of transgender persons.' Each response from ChatGPT-3.5, ChatGPT-4o and Google Gemini were independently evaluated by two endocrinologists [S.T. and S.H.O.] using a 5-point Likert scale, based on adherence to the

Table 1. Compatibility of AI-generated answers to common questions about gender affirming care with current guidelines [5-point likert scale].

	ChatGPT-3.5		ChatGPT-4o		Google Gemini	
	ES	WP	ES	WP	ES	WP
A- Gender Affirming Hormonal Therapy [GAHT]						
What is gender affirming hormonal therapy [GAHT]?	5	5	5	5	5	5
What are the effects of GAHT?	4	4	5	5	3	3
Who are eligible for GAHT?	5	3	5	4	3	3
Which medications are used in GAHT, and how are they administered?	4	4	5	5	2	2
When do the effects of GAHT typically appear?	3	3	4	4	2	2
How long after starting GAHT does menstruation stop?	5	5	5	5	4	4
What additional steps can I take, besides GAHT, to promote facial hair growth?	2	2	2	2	2	2
When does the voice begin to change with GAHT, and how can I modify my voice?	NA	2	NA	5	NA	5
Does GAHT cause weight gain?	5	5	5	5	5	5
Is it possible to have children after starting GAHT?	3	3	4	4	2	2
Is GAHT a lifelong treatment?	NA	NA	NA	NA	NA	NA
Are treatments related to gender transition covered by health insurance?	NA	5	NA	5	NA	5
B-Adverse Outcomes & Long-Term Care						
What are the health risks and potential side effects of GAHT?	4	4	4	4	3	3
Do GAHT increase the risk of cancer?	5	5	5	5	3	3
How often should I be checked for my hormone levels and potential side effects?	5	5	5	5	3	3
Which blood tests are required to monitor GAHT?	5	5	5	5	4	4
C- Gender-Affirming Surgery						
Who are eligible for gender-affirming surgery?	5	5	5	5	3	3
What options are available for gender-affirming surgeries?	3	4	4	5	3	2

ES: Endocrine Society Guideline 2017 for endocrine treatment of gender dysphoric/gender incongruent persons [15], WPATH: World Professional Association for Transgender Health [WPATH] Standards of Care 8 [14], NA: Not available.

Endocrine Society [ES] Guideline 2017 for endocrine treatment of gender dysphoric/gender incongruent persons [15] and World Professional Association for Transgender Health [WPATH] Standards of Care 8 [14] (Figure 1):

1. Fully adheres to the guidelines without deviation
2. Mostly adheres but with minor omissions or deviations
3. Partially adheres with some notable gaps or inconsistencies

4. Minimally adheres and includes significant inaccuracies or gaps

5. Does not adhere to the guidelines or contradicts them

Evaluation of reliability, quality, tendency to hallucinate and readability

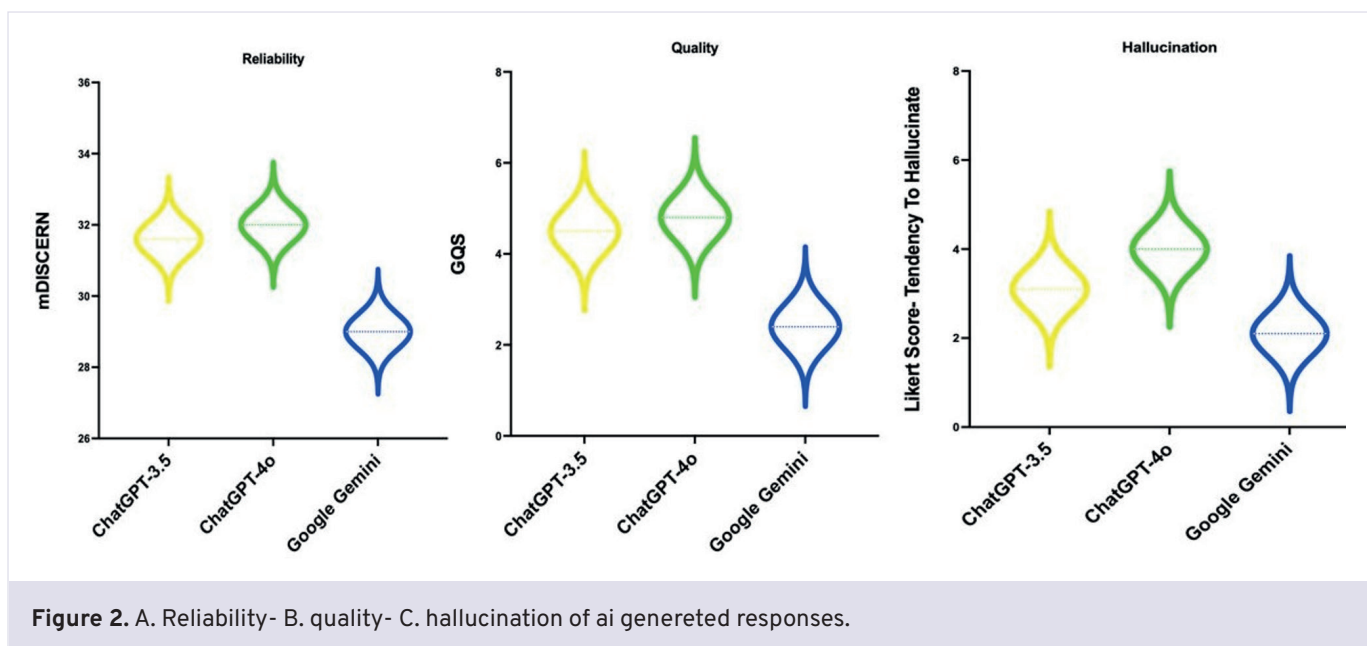
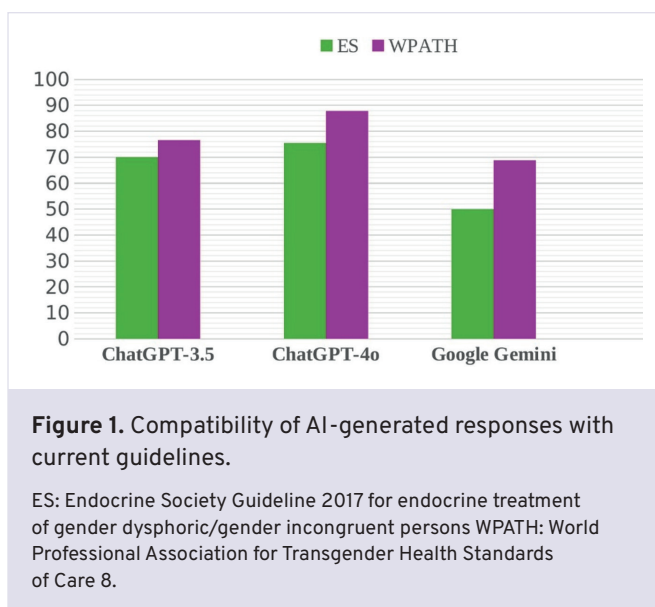
Each response from ChatGPT-3.5, ChatGPT-4o, and Google Gemini was independently evaluated by two endocrinologists. Reliability, quality, and tendency to hallucinate were scored by both evaluators, and a consensus score was determined (Figure 2).

The DISCERN scale, a widely used tool for assessing the reliability and quality of online health information, was employed in this study [16,17]. The scale is divided into three sections: the first section consists of eight questions assessing the reliability of the information, the second section includes seven questions evaluating the quality of information on treatment options, and the final section addresses the overall quality of the publication as a source of treatment information. This study utilized the modified DISCERN [mDISCERN] scale, which focuses exclusively on the first section of the original DISCERN scale. Responses were scored

as follows: a “no” response was given a score of 1, a “partial” response was scored between 2 and 4, and a “yes” response received a score of 5. The total scores were categorized as follows: below 40% [8–15] was considered poor, 40–79% [16–31] was rated as fair, and above 80% [32–40] was deemed good [17].

The Global Quality Scale [GQS], a tool commonly used in related studies, was applied to assess the quality of ChatGPT’s responses [18]. On this scale, a score of 1 indicates poor quality, while a score of 5 represents excellent quality. The GQS also facilitates quality classification, with scores of 1–2 categorized as low quality, 3 as moderate quality, and 4–5 as high quality. ChatGPT has gained recognition as a remarkable global innovation, capable of producing highly realistic texts within seconds. However, it carries the potential risk of spreading inaccurate information and misconceptions, a phenomenon referred to by technical experts as “hallucination” [19]. Tendency to hallucinate of these responses were independently evaluated by two endocrinologists using a 5-point Likert scale.

Readability was assessed using Readable software [Readable.com, Horsham, United Kingdom] [20]. The analysis was conducted based on three widely recognized metrics: the Flesch Reading Ease [FRE] score, the Flesch-Kincaid Grade Level [FKGL], and the Gunning Fog Index [GFI]. An increase in the FRE score and a decrease in the other two show superior readability.



Data analysis

The Kolmogorov-Smirnov test was utilized to evaluate the normality of the data distribution. Numerical variables were presented as mean or median values. The accuracy and quality of the AI models were analyzed using frequencies and percentages and compared using Chi-square tests. For comparisons between chatbots, the mean values of accuracy, quality scores, and readability metrics were calculated and analyzed using the one-way ANOVA method. All statistical analyses were conducted using SPSS version 26.0 [IBM Corp., Armonk, NY, USA]. A p-value of <0.05 was considered the threshold for statistical significance.

Results

A total of 18 questions were posed to ChatGPT-3.5, ChatGPT-4o and Google Gemini regarding gender affirming care. Table 1 presents the compatibility of each answer of ChatGPT-3.5, ChatGPT-4o and Google Gemini with the Endocrine Society Guideline 2017 and WPATH Standards of Care 8 [14,15]. The overall compatibility of the responses from ChatGPT-3.5, ChatGPT-4o, and Google Gemini with the Endocrine Society and WPATH guidelines was 70% and 76.6%, 75.5% and 87.8%, and 50% and 68.8%; respectively. Responses to questions in Category B, which addressed adverse outcomes and long-term care, demonstrated the highest alignment with established guidelines. ChatGPT-3.5 and ChatGPT-4 achieved a 95% concordance with ES and WPATH guidelines, while Google Gemini achieved a 65% concordance with ES and WPATH. This was followed by responses in Category C, addressing gender-affirming surgery, where ChatGPT-4 achieved concordance

rates of 90% with ES and 100% with WPATH guidelines. ChatGPT-3.5 demonstrated 80% and 90% concordance with ES and WPATH guidelines, respectively, while Google Gemini showed lower rates, at 60% for ES and 50% for WPATH. In contrast, responses in Category A, focused on gender-affirming hormone therapy [GAHT], demonstrated the lowest alignment with guidelines. ChatGPT-4 achieved concordance rates of 66.6% with ES and 81.6% with WPATH, while ChatGPT-3.5 showed 60% concordance with both. Google Gemini had the lowest alignment, with 46.6% to ES and 63.3% to WPATH (Table 1). The mean mDISCERN score, representing the “accuracy” of responses, was highest for ChatGPT-4o [32 ± 0.4], followed by ChatGPT-3.5 [31.6 ± 2.4] and Google Gemini [29 ± 0.2] (Table 2). While ChatGPT-4o exhibited comparable accuracy to ChatGPT-3.5 [$p = 0.12$], it significantly outperformed Google Gemini [$p < 0.001$]. ChatGPT-3.5 also demonstrated significantly higher accuracy than Google Gemini [$p < 0.001$]. The mean GQS score, reflecting the “quality” of responses, was highest in ChatGPT-4o [4.8 ± 0.3], followed by ChatGPT-3.5 [4.5 ± 0.5] and Google Gemini [2.4 ± 0.5]. ChatGPT-4o showed similar quality to ChatGPT-3.5 [$p = 0.19$] but significantly outperformed Google Gemini [$p < 0.001$]. Additionally, ChatGPT-3.5 demonstrated higher quality compared to Google Gemini [$p < 0.001$]. On the other hand, mean scores for ‘tendency to hallucinate’ was highest for ChatGPT-4o [4 ± 0.7], followed by ChatGPT-3.5 [3.1 ± 0.3] and Google Gemini [2.1 ± 0.3], [$p < 0.001$], Table 2. Among ‘readability’ indexes, Flesch–Kincaid grade level [FKGL] was highest for ChatGPT-4o [12.6 ± 1.3], followed by ChatGPT-3.5 [12.1 ± 1.8] and Google Gemini [11.8 ± 1.6]. While FKGL did not differ between ChatGPT-4o and ChatGPT-3.5 [$p = 0.95$], ChatGPT-4o generated significantly more

Table 2. The accuracy, quality, tendency to hallucinate, and readability scores of AI-generated answers.

	ChatGPT-3.5	ChatGPT- 4o	Google Gemini
mDISCERN [mean±SD]	31.6±2.4	32 ± 0.4	29±0.2
GQS [mean±SD]	4.5±0.5	4.8±0.3	2.4±0.5
Tendency to hallucinate [mean±SD]	3.1±0.3	4±0.7	2.1±0.3
Readability			
Flesch Kincaid Grade [mean±SD]	12.1±1.8	12.6±1.3	11.8±1.6
Gunning Fox Index [mean±SD]	12.8±1.8	13.1±1.8	13.4±2.4
Flesch Reading Ease [meanSD]	27±3.5	21.4±9.5	32.2±9.6

mDISCERN: modified DISCERN, SD: Standart Deviation, GQS: The Global Quality Scale .

readable responses than Google Gemini [lower FKGL; $p < 0.001$]. Gunning Fog Index [GFI] was similar between Google Gemini [13.4 ± 2.4] and ChatGPT-4o [13.1 ± 1.8] but exceeded that of ChatGPT-3.5 [12.8 ± 1.8] with statistical significance, [$p < 0.001$]. Flesch Reading Ease [FRE] was the highest for Google Gemini [32.2 ± 9.6] followed by ChatGPT-3.5 [27 ± 3.5] and ChatGPT-4o [21.4 ± 9.5] respectively, [$p < 0.001$], Table 2.

Discussion

This study evaluated the efficacy and alignment of responses from ChatGPT-3.5, ChatGPT-4o, and Google Gemini with two current guidelines on the healthcare of transgender people [12,13]. The overall accuracy of the responses from ChatGPT-3.5, ChatGPT-4o, and Google Gemini with ES and WPATH guidelines was 70% and 76.6%, 75.5% and 87.8%, and 50% and 68.8%; respectively. The most notable differences in alignment with the two guidelines were observed in areas such as GAHT eligibility, voice therapy [which is addressed only superficially in the ES guideline], surgical options, and insurance coverage. This discrepancy may simply reflect the fact that the WPATH SOC8 is more up-to-date than the ES guideline. The ES guideline might not have included certain topics due to limited evidence at the time of its publication. Furthermore, the WPATH SOC8 is multidisciplinary and intended for a wide range of healthcare providers, aiming to offer more comprehensive transgender health care. The ES guideline, on the other hand, primarily targets endocrinologists and may prioritize hormonal aspects over adjunct treatments.

Nonetheless, the responses of ChatGPT-3.5 and Google Gemini to the question regarding the timing of GAHT effects were equally misaligned with both guidelines. In contrast to the guidelines, they suggested earlier timetables for skin changes and later timelines for fat redistribution when GAHT was started. Additionally, they included emotional changes, energy shifts, and mood alterations as expected effects of GAHT, which were not extensively covered in the guidelines. ChatGPT-4o distinguished itself by referring to WPATH guidelines and delivering more detailed answers, although some inaccuracies were still noted in its timelines. Nonetheless, all AI systems advised individuals about the importance of GAHT use under the supervision of healthcare professionals, and emphasized the individualized nature of the treatment. Likewise, the answers about the

frequency of hospital visits emphasized that such visits can be tailored to individual needs and circumstances.

Notably, ChatGPT-3.5 and ChatGPT-4o provided non-evidence-based options for facial hair growth and voice therapy, beyond what is included in the guidelines. For facial hair growth, they suggested additional interventions such as micro-needling, exfoliation, supplements, herbal oils, and avoiding shaving, none of which are supported by evidence. For voice training, working with online apps in addition to working with a speech pathologist was offered. They also addressed fertility with optimism, despite limited evidence, while emphasizing the need to discuss fertility preservation before starting GAHT. Google Gemini struggled to provide clear answers to these questions, leading to shorter and less detailed responses. Gemini's responses were more generic, prioritizing consultation with healthcare professionals and referencing sources. On the other hand, it was not possible to assess the LLMs' alignment on the question of whether GAHT is a lifelong treatment, as this issue is not addressed in either guideline. ChatGPT-3.5 and -4o consider GAHT as a lifelong treatment for most transgender persons, depending on individual needs and circumstances.

Furthermore, it was observed that ChatGPT-3.5 and -4o mentioned surgical options beyond what is included in the guidelines, such as hip augmentation and muscle implants, which are not yet sufficiently covered in the literature. Both exhibited general knowledge and emphasized that the timing of GAS is a decision that can be influenced by eligibility, and many other factors such as individual preferences and the overall goals, as well as financial considerations. However, we did not pose other detailed questions about gender-affirming surgery in this study, as recent studies have already thoroughly assessed their responses related to this topic [21,22].

Regarding the accuracy and quality of answers related to gender-affirming care, Google Gemini's total score was significantly lower than that of ChatGPT-4o, while no significant difference was found between the scores of ChatGPT-4o and ChatGPT-3.5, consistent with findings from previous studies on various health topics [23-26]. The answers provided by Google Gemini and ChatGPT-3.5 were slightly better for topics related to adverse outcomes and gender-affirming surgery compared to GAHT. However, both were notably inferior to ChatGPT-4o across all areas. This difference can

be attributed to several factors. ChatGPT-3.5 was trained on data up to January 2022, while ChatGPT-4o, released in March 2023, is an improved version with a better model and more medical data. Additionally, ChatGPT-4o can fine-tune for specific domains and better understand specialized medical terminology in context. More importantly, ChatGPT-4o enhances the accuracy of answering questions in specific medical fields by building on user feedback from ChatGPT-3.5. Google Gemini, developed by Google is fundamentally different from ChatGPT-4o, developed by OpenAI, in how it processes information and generates answers [27-29]. However, consistent with previous studies, ChatGPT-4o demonstrated a higher tendency to produce hallucinations compared to ChatGPT-3.5 and Google Gemini [30,31].

Overall, it was noted that the readability level of all chatbots was fairly low. Based on the scores, only those with at least a high school education were able to understand the responses effectively. We determined that the readability order of AI answers regarding gender affirming care, from easy to difficult, is Gemini, and ChatGPT-3.5 and ChatGPT-4o which is consistent with a recent study [18,32,33].

While this study is among the first to explore the potential of AI in addressing transgender healthcare related answers, several limitations should be acknowledged. The lack of significant differences in response quality between ChatGPT-4 and 3.5 may be attributed to the relatively limited number of questions analyzed. Additionally, the subjective nature of response evaluation should be considered a methodological constraint, as individual interpretations may influence the assessment of accuracy and relevance. Another limitation is the study's inclusion of only two large language models, ChatGPT and Gemini, despite the availability of several alternatives, such as Perplexity.ai, Claude by Anthropic, Microsoft Copilot, Llama2 by Meta, and Mistral.ai. While these models may demonstrate superior performance in general AI applications, their effectiveness in medical and sensitive sociocultural contexts remains limited and largely unverified. The rapid advancement of AI technology indicates that chatbot performance is likely to evolve over time, potentially affecting the long-term applicability and generalizability of the study's findings. Additionally, despite the simultaneous administration of identical questions to the models, response variability at different time points remains a concern, as it may

be influenced by updates to training datasets and modifications to model parameters.

In conclusion, while ChatGPT-4o demonstrated superior alignment with guidelines and provided more detailed responses compared to ChatGPT-3.5 and Google Gemini in the context of gender-affirming care, its tendency to produce hallucinations underscores the need for cautious application in medical contexts. Although AI models like ChatGPT and Google Gemini can serve as valuable tools for general medical information, they are not substitutes for professional medical advice. Patients should remain cautious when utilizing these AI models, as their responses may include inaccuracies or lack the specificity needed for individual situations. The variations in readability and accuracy among these AI models highlight the importance of continuous evaluation and refinement to ensure that they deliver accessible, reliable, and evidence-based information. This study emphasizes the necessity for healthcare professionals to verify AI-generated content and integrate these tools to support, rather than replace, clinical expertise. Therefore, patients are strongly advised to consult qualified healthcare providers to make informed decisions regarding their health and treatment options.

Author contributions

Conception: S.T., B.E., S.H.O., B.O.Y.; Design: S.T., S.H.O., B.O.Y.; Data acquisition: S.T., B.E.; Data analysis: S.T., B.E.; Data interpretation: S.T., B.E.; Drafting of the manuscript: S.T.; Critical revision of the manuscript: S.T., S.H.O., B.O.Y. All authors reviewed the results, approved the final version of the manuscript, and agreed to be accountable for all aspects of this study.

Ethical approval

Ethics committee approval and informed consent were not required for this study.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest

The authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The authors declare that this study received no funding.

Generative AI statement

The authors declare that no generative AI or AI-assisted technologies were used in the writing or preparation of this study.

References

- [1] Spivey LA, Edwards-Leeper L. Future directions in affirmative psychological interventions with transgender children and adolescents. *J Clin Child Adolesc Psychol* 2019;48(2):343-56. [\[Crossref\]](#)
- [2] Winter S, Diamond M, Green J, et al. Transgender people: health at the margins of society. *Lancet* 2016;388(10042):390-400. [\[Crossref\]](#)
- [3] Call DC, Challa M, Telingator CJ. Providing affirmative care to transgender and gender diverse youth: disparities, interventions, and outcomes. *Curr Psychiatry Rep* 2021;23(6):33. [\[Crossref\]](#)
- [4] Korpaisarn S, Safer JD. Gaps in transgender medical education among healthcare providers: a major barrier to care for transgender persons. *Rev Endocr Metab Disord* 2018;19(3):271-5. [\[Crossref\]](#)
- [5] Teti M, Kerr S, Bauerband LA, Koegler E, Graves R. A Qualitative scoping review of transgender and gender non-conforming people's physical healthcare experiences and needs. *Front Public Health* 2021;9:598455. [\[Crossref\]](#)
- [6] Heng A, Heal C, Banks J, Preston R. Transgender peoples' experiences and perspectives about general healthcare: a systematic review. *International Journal of Transgenderism* 2018;19(4):359-78. [\[Crossref\]](#)
- [7] OpenAI. Introducing ChatGPT. Available at: <https://openai.com/index/chatgpt/>
- [8] Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds and Machines* 2020;30:681-94. [\[Crossref\]](#)
- [9] Google DeepMind. Gemini. Available at: <https://deepmind.google/technologies/gemini/#introduction>
- [10] Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using chatgpt in health care. *J Med Internet Res* 2023;25:e48009. [\[Crossref\]](#)
- [11] Criss S, Nguyen TT, Gonzales SM, et al. "HIV Stigma Exists" - Exploring ChatGPT's HIV advice by race and ethnicity, sexual orientation, and gender identity. *J Racial Ethn Health Disparities* 2025;12(6):3622-35. [\[Crossref\]](#)
- [12] Link BG, Phelan JC. Conceptualizing stigma. *Annual Review of Sociology* 2001;27(1):363-85. [\[Crossref\]](#)
- [13] Velasco RAF, Slusser K, Coats H. Stigma and healthcare access among transgender and gender-diverse people: a qualitative meta-synthesis. *J Adv Nurs* 2022;78(10):3083-100. [\[Crossref\]](#)
- [14] East Dundee I, Tax I, Vella B. World professional association for transgender health [WPATH].
- [15] Hembree WC, Cohen-Kettenis PT, Gooren L, et al. Endocrine treatment of gender-dysphoric/gender-incongruent persons: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab* 2017;102(11):3869-903. [\[Crossref\]](#)
- [16] Ozduran E, Büyükçoban S. Evaluating the readability, quality and reliability of online patient education materials on post-covid pain. *PeerJ* 2022;10:e13686. [\[Crossref\]](#)
- [17] Kumar VS, Subramani S, Veerapan S, Khan SA. Evaluation of online health information on clubfoot using the DISCERN tool. *J Pediatr Orthop B* 2014;23(2):135-8. [\[Crossref\]](#)
- [18] Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep* 2024;14(1):243. [\[Crossref\]](#)
- [19] Ahmad Z, Kaiser W, Rahim S. Hallucinations in ChatGPT: an unreliable tool for learning. *Rupkatha J Interdiscip Stud Humanit* 2023;15(4):12. [\[Crossref\]](#)
- [20] Readability Formulas. Available at: <https://readabilityformulas.com/RcRsRtRlCrf>
- [21] Najafali D, Hinson C, Camacho JM, Galbraith LG, Tople TL, Eble D, et al. Artificial intelligence knowledge of evidence-based recommendations in gender affirmation surgery and gender identity: is ChatGPT aware of WPATH recommendations? *Eur J Plast Surg* 2023;46(6):1169-76. [\[Crossref\]](#)
- [22] Snee I, Lava CX, Li KR, Corral GD. The utility of ChatGPT in gender-affirming mastectomy education. *J Plast Reconstr Aesthet Surg* 2024;99:432-5. [\[Crossref\]](#)
- [23] Cinar C. Analyzing the performance of ChatGPT about osteoporosis. *Cureus*. 2023;15(9). [\[Crossref\]](#)
- [24] Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. *J Orthop Surg Res* 2024;19(1):574. [\[Crossref\]](#)

- [25] Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-Generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina* 2023;7(10):862-8. [\[Crossref\]](#)
- [26] Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* 2023;95. [\[Crossref\]](#)
- [27] Mihalache A, Grad J, Patil NS, et al. Google Gemini and bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. *Eye (Lond)* 2024;38(13):2530-5. [\[Crossref\]](#)
- [28] Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. *Eye (Lond)* 2024;38(8):1412-7. [\[Crossref\]](#)
- [29] Rane N, Choudhary S, Rane J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *SSRN Electronic Journal* 2024. [\[Crossref\]](#)
- [30] Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform* 2024;12. [\[Crossref\]](#)
- [31] Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 2023;15(2):e35179. [\[Crossref\]](#)
- [32] Reyhan AH, Mutaf Ç, Uzun İ, Yüksekayla F. A Performance evaluation of large language models in keratoconus: a comparative study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. *J Clin Med* 2024;13(21):6512. [\[Crossref\]](#)
- [33] McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an artificial intelligence chatbot for delivery of ir patient education material: a comparison with societal website content. *J Vasc Interv Radiol* 2023;34(10):1760-8 [\[Crossref\]](#)