ORIGINAL ARTICLE

Eastern Journal of Medicine
East J Med 31(1): 1-5, 2026
DOI: 10.5505/ejm.2026.77861

Comparative Analysis of the Accuracy and Readability of the Answers Given by ChatGPT, DeepSeek and Gemini Artificial Intelligences about Penile Body Dysmorphic Disorder

Murat Demir[1*], Muhammed Kotan[1], Muhammed Hamdi Sarihan[1], Kasım Ertaş[2], Rahmi Aslan[1], Recep Eryılmaz[1], Kerem Taken[1]

[1]*Van Yuzuncu Yil University,* Faculty of Medicine, Department of Urology, Van
[2]*Private Lokman Hekim Hospital, Urology, Van*

ABSTRACT

This study contrasts the validity and legibility of responses of three computer algorithms—ChatGPT, DeepSeek, and Gemini to Penile Body Dysmorphic Disorder (BDD), where individuals are concerned with penis size. As individuals more and more rely on AI to seek health knowledge, this research contrasts the validity and readability of AI-generated responses to common penile size and function questions.

Fourteen frequently asked questions about penis size and function were collected from social media and posed to ChatGPT, DeepSeek, and Gemini. Responses were evaluated with the Discern index to evaluate reliability and the Gunning Fog index to measure readability. For statistical analysis one-way ANOVA and one-way ANOVA and Post hoc tests were used.

There was no significant difference between three algorithms (p=0.063), and the distinct scores for ChatGPT, DeepSeek, and Gemini were 3.29±1.07, 3.86±0.66, and 3.43±0.85. In comparison to DeepSeek (19.81±1.88) and Gemini (20.07±2.37), ChatGPT produced a significantly lower Gunning Fog score of 18.34±1.72 (p=0.012), suggesting that ChatGPT responses were easier to read.

The study found all three AI models provided accurate and informative content about Penile BDD, where DeepSeek provided the most precise and ChatGPT the easiest to read. However, AI should never replace expert medical recommendations. Continuous improvement and ethical reasoning are required to ensure safe application of AI for health information.

Keywords: Penis, BDD, ChatGBT

Introduction

Men's penis size is usually thought to be a gauge of their sexual performance and masculinity. Cultural and societal expectations may reinforce the idea that penis size matters and thus lead men to be concerned about how their penis will be viewed during sex. Such concern may result in sexual dysfunction. Excessive preoccupation with having a normally sized penis is known as "small penis anxiety" or "small penis syndrome" in the literature. Patients with micropenis are not included in this definition. A diagnosis of Body Dysmorphic Disorder (BDD) may fulfill the criteria for some men with excessive worry and embarrassment regarding the size of their penis (1).

BDD is a psychiatric illness in which one is overly fixated on physical imperfections that are too small or non-existent to be noticed by others. The individual with BDD may form repetitive behaviors like constantly checking these perceived defects and comparing them with others and may experience extreme distress or impairment of social, occupational or other important functional areas (2). Suicidal thoughts and completed suicide rates have been said to be elevated among individuals with BDD in psychiatric settings. BDD

*Corresponding Author: Assist. Prof. Dr. Murat Demir, Department of Urology, Faculty of Medicine, Van Yuzuncu Yıl University, 65080-Van, Turkey
E-mail: urologmurat72@gmail.com, Phone: +90 432 225 17 01-05, Fax: +90 432 236 1054
ORCID ID: Murat Demir: 0000-0001-5029-8800, **Muhammed Kotan:** 0009-0009-3504-3725, **Muhammed Hamdi Sarihan:** 0009-0006-7759-9201, **Kasım Ertaş:** 0000-0003-4300-1399, **Rahmi Aslan:** 0000-0002-4563-0386, **Recep Eryılmaz:** 0000-0002-4506-8784, **Kerem Taken:** 0000-0002-4370-4222
Received: 25.03.2025, **Accepted:** 07.01.2026

usually starts in adolescence and may have a chronic course until appropriate treatment is started (2).

Current studies report that men experiencing body dysmorphic concerns regarding penis size experience more shame and are seriously compromised in their relationships (3). Penis size concerns negatively affect general mental health and sexual functioning in men. These people always check the size of their penis, compare with others and even attempt to change this situation with surgery or cosmetic interventions (3,4). However, as penile enlargement surgery or other cosmetic procedures lack scientific background and sound results, whether these therapies are effective or not is arguable.

Today, many men look for online sources to talk about their penis size and sexual functioning issues and use AI-assisted systems to retrieve information. There are AI systems such as ChatGPT, DeepSeek, Gemini that educate users on various medical and psychological issues. Very little research is conducted on the validity and scientific basis of information provided through these systems. Throughout the current research study, effectiveness of the information provided by ChatGPT, DeepSeek, Gemini in terms of size, thickness, and performance in the case of persons with penile dysmorphic disorder or fixation on small penis will be evaluated and trustfulness in AI-informants in such persons will be examined.

Material and Method

In this study, the 14 most frequently asked penis size queries on social media were identified. These queries were posed to ChatGPT, DeepSeek and Gemini AI models and the response of each model was recorded. The answers were rated by the Discern and Gunning Fog measures to investigate readability and content quality.

Ethics committee approval was not obtained because patient data was not used.

1. **Data Collection Process:** The 14 most frequently asked penis size and function questions on social media were ascertained.

The questions were posed to ChatGPT, DeepSeek and Gemini AI models respectively.

All three models' answers were recorded and stored for analysis.

2. **Evaluation Criteria:** The scientific validity and reliability of the responses were measured by the Discern index. The Discern index is a tool used to assess the quality and reliability of health information content. It is a 6-item tool that tests health-related written and electronic sources on independence, sourcing and impartiality. The better the Discern score, the more reliable the content (5).

The intelligibility and readability of the responses were tested using the Gunning Fog index. The index takes into account sentence length and word complexity to determine the readability level of a text. The higher the Gunning Fog, the more difficult it is to comprehend the text. Scholarly and scientific writing tends to have higher scores in Gunning Fog than texts written for the general public.

Data Analysis: Descriptive statistics for continuous variable are given as Mean± Standard Deviation. One-way ANOVA was performed for the comparison of group means. Duncan multiple comparison test was also used to identify different groups. Statistical significance level was considered as 5% in calculations and analyses were done through SPSS (Version: 21, IBM Corporation, Armonk, NY) statistical package software.

Results

In our study, the Discern scores were ChatGPT 3.29± 1.07, DeepSeeK 3.86± 0.66 and Gemini 3.43± 0.85 (p=0.063). No difference between the DISCERN scores of within-group models from each other was found (Table 1). In a comparison of the Gunning Fog scores, ChatGPT 18.34± was 1.72, DeepSeeK 19.81± was 1.88, Gemini 20.07± was 2.37 (p 0.012). Comparing in-group models with each other, ChatGPT showed a significantly lower Gunning Fog measure compared to DeepSeek and Gemini (Table 2).

Table 1 Comparison of DISCERN scores among artificial intelligence models.

No statistically significant differences were found between the AI models in terms of their DISCERN scores ($p > 0.05$). Although ChatGPT showed slightly higher mean values compared with DeepSeek ($p = 0.098$) and Gemini ($p = 0.816$), these differences were not significant. Similarly, the difference between DeepSeek and Gemini ($p = 0.227$) was not statistically significant, indicating comparable overall performance across the models.

Table 2 Comparison of Gunning Fog scores among artificial intelligence models.

Statistically significant differences were observed

Table 1: Comparison of DISCERN Scores of Artificial Intelligence Models

Artificial Intelligence Models	P Value
ChatGPT vs. DeepSeek	0.098
ChatGPT vs. Gemini	0.816
DeepSeek vs. Gemini	0.227

Table 2: Comparison of Gunning Fog scores of Artificial Intelligence Models

Artificial Intelligence Models	P Value
ChatGPT vs. DeepSeek	0.042
ChatGPT vs. Gemini	0.008
DeepSeek vs. Gemini	0.623

between the models in terms of their Gunning Fog scores. ChatGPT demonstrated significantly higher readability complexity compared with both DeepSeek ($p = 0.042$) and Gemini ($p = 0.008$). However, there was no significant difference between DeepSeek and Gemini ($p = 0.623$). These findings suggest that ChatGPT's generated texts tend to be more complex and less readable than those of the other models.

Discussion

This research investigated the performance of three popular large language models (LLMs)—ChatGPT, DeepSeek, and Gemini—in responding to questions about body dysmorphic disorder (BDD), a condition frequently observed among young adults. The results indicate that these LLMs are capable of delivering accurate and insightful answers regarding BDD. Specifically, ChatGPT offered detailed and thorough responses, whereas DeepSeek provided clearer and more accessible information. Gemini struck a balance between the two, offering both informative and easy-to-understand answers.

BDD is increasingly prevalent among young people, making it crucial for LLMs to provide information that is straightforward, and easy to grasp. BDD is a mental health disorder characterized by individuals fixating on perceived flaws, often leading to significant psychological distress. As such, AI-powered systems have the potential to play a vital role in disseminating accurate information about BDD to young audiences, helping them recognize the issue and take appropriate action when needed (7, 8).

Although the findings demonstrate that LLMs can deliver highly accurate and useful responses to BDD-related queries, certain limitations remain. The automated outputs generated by LLMs cannot substitute for professional medical care. Conditions like BDD require the expertise and intervention of trained healthcare providers for effective treatment (9).

Moreover, it is important to recognize that cultural differences significantly influence how individuals perceive both body image and AI-generated health information. Attitudes toward body appearance, mental health, and technology vary widely across societies, which may shape how users interpret responses about BDD. For instance, in some cultures, discussing body dissatisfaction or mental health may still be stigmatized, while in others, open conversations are more common. Consequently, LLMs should be culturally adaptive, ensuring that their language, tone, and examples are sensitive to diverse social norms and values. Incorporating cultural context into LLM training could enhance the relevance and acceptance of AI responses worldwide, making the information about BDD more inclusive and effective for users from different backgrounds (10).

In summary, the reliability of LLMs' responses depends not only on the sophistication of their algorithms but also on the quality of the data they are trained on. To ensure the accuracy and consistency of LLMs' answers regarding BDD, it is essential to regularly update and validate their knowledge bases with the latest research findings and enhance their training to improve precision (11).

Specifically, it should be stressed that the responses of LLMs to BDD questions should avoid persuading teenagers to diagnose BDD symptoms themselves. Besides, chatbots are to help teenagers reach to a qualified person by providing them with a link to a reliable source of information. Not only by telling the true information about BDD but also by guiding the

children to the tools, which can help them to achieve the desired body image are suggested.

Furthermore, the balance between readability and reliability in LLM-generated responses plays a crucial role in shaping users' decision-making processes, particularly among young individuals seeking health-related information. While easily understandable content can enhance engagement and comprehension, oversimplification may risk omitting essential clinical nuances, potentially leading to misinterpretation or self-diagnosis. Conversely, overly technical explanations, though accurate, might alienate non-expert users and reduce the accessibility of vital information. Therefore, achieving an optimal balance between readability and reliability is essential to ensure that users can make informed and responsible health decisions. This balance can help individuals recognize when professional consultation is needed, promoting safe and ethical use of AI-assisted health information systems (12,13).

Thus, it is suggested that parents should teach adolescents to associate synthase ever modification with growth and maturation, in that way, allowing teenagers to comprehend messages more clearly and also learn from the LMs more effectively.

In other words, the study at hand exhibits that LLMs can improve BDD recognition in the youth and communicate the accurate points through them.

Limitations and Future Studies: This research has some limitations. First, the performance of AI models as tested in their replies in this study is fixed at a specific time (February 2025). With the constant change of AI technologies, the performance of these AI models in the future will look different.

Second, our research only assessed three AI models. There are many AI models on the market and they may have various performances (14).

Larger studies might be conducted to assess a wider range of AI models and the role of such models on different populations (e.g., different ages, genders, cultures). The article stresses the necessity of AI-driven health information websites being an early tool in the detection and treatment of BDD a mental illness.

Lastly, this study confirms that artificial intelligence models such as ChatGPT, Deepseek and Gemini can be in a position to provide accurate and useful information to adolescents on such sensitive issues like penis size. These models are however required to be designed, revised and utilized on a consistent basis in terms of ethics. Health information systems based on artificial intelligence may be a potent weapon against such mental disorders as BDD, but such tools must be used responsibly and with caution.

References

1. Veale D, Miles S, Read J, Troglia A, Wylie K, Muir G. Sexual functioning and behavior of men with body dysmorphic disorder concerning penis size compared with men anxious about penis size and with controls: A cohort study. Sex Med 2015; 3: 147-155.
2. Laughter MR, Anderson JB, Maymone MB, Kroumpouzos G. Psychology of aesthetics: Beauty, social media, and body dysmorphic disorder. Clin Dermatol 2023; 41: 28-32.
3. Fabris MA, Longobardi C, Badenes-Ribera L, Settanni M. Prevalence and co-occurrence of different types of body dysmorphic disorder among men having sex with men. J Homosex 2022; 69: 132-144.
4. Mansfield AK. Genital manifestations of body dysmorphic disorder in men: A review. Fertil Steril 2020; 113: 16-20.
5. Rees CE, Ford JE, Sheard CE. Evaluating the reliability of DISCERN: A tool for assessing the quality of written patient information on treatment choices. Patient Educ Couns 2002; 47: 273-275.
6. Świeczkowski D, Kułacz S. The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of health literacy challenges in medical linguistics: An exploratory study. Cardiol J 2021; 28: 627-631.
7. Fang A, Sawyer AT, Aderka IM, Hofmann SG. Psychological treatment of social anxiety disorder improves body dysmorphic concerns. J Anxiety Disord 2013; 27: 684-691.
8. Hasirci E, Kazaz IO, Alkis O, Eker A, Gultekin MH, Kayra MV, Akbaba KT, Duran MB, Ongun S, Toprak T, Ceker G, Deliktas H, Cicek T, Ceyhan E, Akgun U, Colak MY, Turunc T. Evaluation of average penis length and the thoughts of Turkish men and partners. Is surgery necessary? Rev Int Androl 2024; 22: 74-81.
9. Ryding FC, Kuss DJ. The use of social networking sites, body image dissatisfaction, and body dysmorphic disorder: A systematic review of psychological research. Psychol Pop Media 2020; 9: 412-426.
10. Abdoli M, Scotto Rosato M, Desousa A, Cotrufo P. Cultural differences in body image: A systematic review. Social Sciences. 2024; 13: 305.

11. Johnson KW, Torres Soto J, Glicksberg BS, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018; 71: 2668-2679.
12. McInnes N, Haglund BJ. Readability of online health information: Implications for health literacy. Informatics for Health and Social Care. 2011; 36: 173-189.
13. Eysenbach G, Köhler C. How do consumers search for and appraise health information on the World Wide Web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002; 324: 573-577.
14. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020; 395: 1579-1586.