Original Investigation | Özgün Araştırma

Gazi Med J 2025;36(4):407-416

DOI: http://dx.doi.org/10.12996/gmj.2025.4451

Comparative Evaluation of Large Language Models in Addressing Autism-Related Information Queries: Insights from ChatGPT, Gemini, and Copilot

Otizm ile İlişkili Soruları Yanıtlamada Büyük Dil Modellerinin Karşılaştırmalı Değerlendirilmesi: ChatGPT, Gemini ve Copilot'tan Elde Edilen Bulgular

Gamze Demir[1], Mehmet Sevri[2], Cafer Doğan Hacıosmanoğlu[3], Dicle Büyüktaşkın[1], Ahmet Özaslan[1,4,5]

[1]Department of Child and Adolescent Psychiatry, Gazi University Faculty of Medicine, Ankara, Türkiye
[2]Department of Computer Engineering, Recep Tayyip Erdoğan University Faculty of Engineering and Architecture, Rize, Türkiye
[3]Department of Child and Adolescent Psychiatry, Yıldırım Beyazıt Univesity Yenimahalle Training and Research Hospital, Ankara, Türkiye
[4]Child Protection Research and Application Center, Gazi University, Ankara, Türkiye
[5]Psychology Research Centre, Khazar University, Baku, Azerbaijan

ABSTRACT

Objective: While large language models (LLMs) have been increasingly evaluated for medical inquiries, their responses to questions about autism spectrum disorder (ASD) remain underexplored. This study aims to evaluate and compare four publicly available LLMs-ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Microsoft Copilot-regarding autism-related queries.

Methods: Nineteen frequently asked autism-related questions categorized into symptoms, diagnosis, treatment, and general information. The responses from each LLM were evaluated by three child and adolescent psychiatrists using the patient education materials assessment tool and the Global Quality Score. Thematic analysis was conducted to identify key topics. A majority consensus approach determined the final ratings, and sentiment analysis was performed to assess emotional polarity and subjectivity.

Results: ChatGPT-4.0 demonstrated superior overall response quality compared to Microsoft Copilot and Google Gemini (p=0.006, p=0.009). While the overall understandability of responses was similar across all LLMs, ChatGPT-4.0 scored significantly higher than Microsoft Copilot on the content subscale (p=0.026), and Google Gemini outperformed ChatGPT-4.0 in word choice and style (p=0.041). Thematic analysis revealed that all chatbots emphasized early diagnosis and behavioral issues. Sentiment analysis indicated a high degree of objectivity across all models. Google Gemini displayed the highest polarity score (0.115),

ÖZ

Amaç: Büyük dil modellerinin (large language models, LLM'ler) tıbbi sorulara verdikleri yanıtlar gün geçtikçe daha fazla araştırılmaktadır; ancak, bu modellerin otizm spektrum bozukluğu (OSB) ile ilgili sorulara verdikleri yanıtlar literatürde yeterince incelenmemiştir. Bu çalışma, otizmle ilişkili sorulara verdikleri yanıtlar açısından dört genel erişime açık LLM'i — ChatGPT-3.5, ChatGPT-4.0, Google Gemini ve Microsoft Copilot — değerlendirmeyi ve karşılaştırmayı amaçlamaktadır.

Yöntemler: Otizmle ilişkili sık sorulan on dokuz soru; belirtiler, tanı, tedavi ve genel bilgi olmak üzere dört kategoriye ayrılmıştır. Her bir LLM'nin yanıtları, üç çocuk ve ergen psikiyatristi tarafından Hasta Eğitimi Materyalleri Değerlendirme Aracı ve Küresel Kalite Skoru kullanılarak değerlendirilmiştir. Tematik analiz ile temel konular belirlenmiş; çoğunluk görüşü yaklaşımıyla nihai puanlar oluşturulmuştur. Duygu analizi, yanıtların duygusal kutupluluğunu ve öznellik düzeyini incelemek amacıyla gerçekleştirilmiştir.

Bulgular: ChatGPT-4.0, genel yanıt kalitesi açısından Microsoft Copilot ve Google Gemini'ye kıyasla üstün performans göstermiştir (p=0,006, p=0,009). Yanıtların genel anlaşılırlığı tüm modeller arasında benzer bulunmakla birlikte, ChatGPT-4.0 içerik alt ölçeğinde Microsoft Copilot'tan anlamlı derecede yüksek puan almıştır (p=0,026). Buna karşılık, Google Gemini kelime seçimi ve üslup açısından ChatGPT-4.0'dan daha iyi performans göstermiştir (p=0,041). Tematik analiz büyük dil modellerinin erken tanı ve davranışsal sorunlara vurgu

Cite this article as: Demir G, Sevri M, Hacıosmanoğlu CD, Büyüktaşkın D, Özaslan A. Comparative evaluation of large language models in addressing autism-related information queries: insights from ChatGPT, Gemini, and Copilot. Gazi Med J. 2025;36(4):407-416

Address for Correspondence/Yazışma Adresi: Gamze Demir, MD, Department of Child and Adolescent Psychiatry, Gazi University Faculty of Medicine, Ankara, Türkiye
E-mail / E-posta: gamzedemir@gazi.edu.tr
ORCID ID: orcid.org/0000-0001-6896-6897

Received/Geliş Tarihi: 08.05.2025
Accepted/Kabul Tarihi: 29.08.2025
Publication Date/Yayınlanma Tarihi: 13.10.2025

while subjectivity scores were moderately high across all chatbots, with ChatGPT-4.0 exhibiting the highest subjectivity score (0.452).

Conclusion: This study highlights the potential of LLMs, particularly ChatGPT-4.0, to deliver high-quality and easily understandable information regarding ASD. However, given the limitations of LLMs, including their susceptibility to biases and lack of real-world reasoning, further research is needed.

Keywords: Autism spectrum disorder, large language models, artificial intelligence, ChatGPT, Gemini, Copilot

yaptığını ortaya koymuştur. Duygu analizi sonuçları, tüm modellerde yüksek düzeyde nesnellik sergilendiğini göstermiştir. Google Gemini en yüksek kutupluluk skoruna (0,115) sahipken, öznellik puanları tüm modellerde orta-yüksek düzeyde bulunmuş, ChatGPT-4.0 en yüksek öznellik skorunu (0,452) göstermiştir.

Sonuç: Bu çalışma, özellikle ChatGPT-4.0'ın, OSB hakkında yüksek kaliteli ve kolay anlaşılabilir bilgiler sunma potansiyeline sahip olduğunu ortaya koymaktadır. Bununla birlikte, LLM'lerin önyargılara yatkın oluşu ve gerçek hayata uygun akıl yürütme eksikliği gibi sınırlılıkları göz önüne alındığında, bu alanda daha fazla araştırmaya ihtiyaç vardır.

Anahtar Sözcükler: Otizm spektrum bozukluğu, büyük dil modelleri, yapay zekâ, ChatGPT, Gemini, Copilot

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by persistent deficits in social communication and interaction and restrictive, repetitive patterns of behavior, activity or interests (1). Receiving the diagnosis of ASD can have negative effects on the entire family system, which includes the need for new skills in adjustment, coping, advocacy, and seeking services for the child (2).

After a diagnosis, parents may experience different stages of emotions, such as shock, fear, grief, and guilt (3). These emotional responses and complexities of the process highlight the considerable challenges faced by parents in caring for individuals with ASD.

Recent years have seen a marked rise in the prevalence of ASD. For example, between 2016 and 2020, prevalence rates rose from one in fifty-four children aged eight years to one in thirty-six (4). This growing prevalence has created a greater demand for mental health services. But there is an inadequate number of qualified mental health professionals and limited infrastructure, leaving families with relatively few available resources (5). As a result, many parents focus on finding solutions and seeking help for their children diagnosed with autism (6).

The process of looking for information on ASD and its management is often described as challenging by parents (7). Several studies suggest that when looking for information and support, parents first seek help from their own social networks (8). Nevertheless, despite these efforts, a variety of parents have noted stigma from their neighboring communities, or have even blamed themselves and other relatives for their child's behavior (9). Such social dynamics can weaken broader support networks and increase feelings of isolation (10,11). This dynamic also makes it even more difficult to access medical support for autism, and pushes caregivers toward alternative ways of gaining information.

The internet has increasingly become a vital resource for families seeking information about ASD and other neurodevelopmental conditions. This trend correlates with the rising prevalence of ASD and the increasing number of online platforms such as social media and online communities (2,12). Online information can be of variable reliability and accuracy, raising questions about the overall quality control of available resources (13). This has increased the need for several online platforms to share trustworthy and easily accessible information with parents (14).

In recent years, machine learning and artificial intelligence (AI) have been increasingly integrated into many aspects of daily life. This integration has made significant progress in the healthcare sector, evidenced by the use of chatbots to facilitate easy access to medical information for individuals (15). Large language models (LLMs) have advanced significantly from conventional natural language processing (NLP) models, introducing innovative capabilities in healthcare services. One of the most popular examples of LLMs is ChatGPT. GPT has evolved considerably since it was introduced in 2018, with the latest model as of March 2023 being GPT4. In addition to ChatGPT, other AI-powered chatbots, such as Google Bard and Microsoft Copilot, are also integrated into several services (16).

LLMs have orders of magnitude more parameters than earlier models. Combined with self-supervised learning on vast datasets, this enables models to generate more human-like responses. These models have introduced innovative approaches to addressing medical inquiries, facilitating computer-aided diagnosis, recommending treatment, and providing health education (16-19). Moreover, they have the ability to inform patients about any health-related issues, answer inquiries relating to health maintenance and disease prevention, as well as provide insights into how social and environmental determinants affect an individual's own health (20,21).

A recent study found that almost 80% of participants (n=607) considered using ChatGPT for self-diagnosis (22). This implies that people turn to chatbots such as ChatGPT to learn more in the health domain, especially symptoms, diagnosis, and treatment. Given that many parents are often not well equipped with knowledge and experience in dealing with ASD, these chatbots serve as a useful and easily available source of information.

Unfortunately, the information obtained from these technologies is not perfect. LLMs have a diversity of major shortcomings, including biases in the training data, the ability to produce disinformation, and a lack of true reasoning capabilities (23,24). A recent study indicated that while ChatGPT demonstrated potential regarding accuracy, comprehensiveness, and speed in clinical psychiatry, it also revealed shortcomings in pharmaceutical information. The shortfall was attributed to ChatGPT's training being predominantly based on web-based information rather than textbooks in the field (25). Hence, it is vital to understand both the pros and cons of these technologies to ensure that they are used efficiently and reliably. For this reason, there is a need for academic research that evaluates the quality of the information provided by such technologies.

The use of LLMs in the healthcare emphasizes how important it is for these systems to respond in a language that is not just clear and understandable, but also non-stigmatizing, empathetic and human-

like the one used by health providers. Previous research examining chatbot responses to health-related questions has shown that they can exhibit empathy and provide accurate answers (26). However, other studies indicate that even LLMs with advanced NLP capabilities may not completely and accurately represent empathy (27,28).

In another study, Spallek et al. (24) examined the accessibility, impartiality, and potential presence of stigmatizing or incorrect language in the outputs of ChatGPT-4. The findings indicated that while the first outputs of ChatGPT-4 were commendable and potentially practical, they still exhibited certain accessibility issues, occasionally employed stigmatizing language, and lacked a diverse array of supportive evidence. These results point to the dangers of LLMs in language use. Accordingly, there is a risk of misguiding or stigmatizing individuals if the language is incorrect or insensitive. These risks are particularly exacerbated in the case of autism-related questions as language has a critical role in shaping attitudes and beliefs about autism (29). Recent studies by the autism research community emphasize the importance of language use in influencing public understanding of autism and related risks (30). Hence, the choice of terminology to characterize autism, particularly the language favored when discussing autistic persons, is crucial in shaping definitions, attitudes, and stigma (31). Consequently, a thorough qualitative assessment of the manner in which AI chatbots handle nuanced and sensitive language when formulating responses to questions about autism is essential for their effective use.

Over the past years, studies have been published on LLMs and their responses to common questions asked in a variety of medical disciplines including cirrhosis, dementia, migraines, uro-oncology, head and neck surgery and vision disorders (14,32-35). However, the exploration of LLM responses to caregivers' frequently asked questions about individuals with neurodevelopmental disorders is not well documented. In this area, McFayden et al. (36) conducted a study assessing the quality of responses given by ChatGPT-4, a widely used AI chatbot, to questions related to ASD. In general, the study showed that ChatGPT-4 was able to generate accurate, concise and easy-to-understand content. However, the study also highlighted areas for improvement, especially with respect to the actionability of the knowledge gained.

To interpret these findings accurately and generalize further, there is a need for research evaluating how well AI systems can answer autism-related frequently asked questions. Investigation of the effectiveness of other AI chatbots like Google's Gemini or Microsoft's Copilot in answering autism-related questions could also help bridge the gap in the literature.

In this study, we evaluate and compare four publicly available LLMs: OpenAI's ChatGPT-3.5 and GPT-4.0 models, Google's Gemini, and Microsoft's Copilot on frequently asked autism-related questions. We then rate each chatbot's responses on understandability and quality based on previously established standards. Furthermore, we conduct qualitative analyses to assess the thematic nature and emotional polarity of the responses generated by the chatbots. Results of these analyses would help understand potential advantages and disadvantages of using AI-powered tools to answer autism-related questions. The study is expected to pave the way for a more comprehensive understanding of how health communication shapes the role and impact of AI-powered chatbots.

MATERIALS AND METHODS

Procedure

The question database was created from informational materials published by organizations such as the American Academy of Child and Adolescent Psychiatry (AACAP), the International Association for Child and Adolescent Psychiatry and Allied Professions, and the European Society for Child and Adolescent Psychiatry. To ensure representation of public and patient concerns, frequently asked questions about ASD from Google Trends were also added to the database. Questions that were repetitive or did not contain medical information were excluded from the study. The questions were categorized into four topics: symptoms, diagnosis, treatment, and general information (Table 1). Grammar corrections were made to ensure clarity and readability. The 19 questions created were directed to ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Microsoft Copilot in English on April 2, 2024, World Autism Awareness Day. The responses were collected and analyzed using new accounts with no previous activity. If parents did not ask the same question twice, a response was requested for each question only once. Each response was independently rated by three child and adolescent psychiatrists with clinical experience in ASD using the Global Quality Score (GQS) and patient education material assessment tool (PEMAT). The medical accuracy of the responses was evaluated according to the AACAP guidelines. As there were no patients involved in the study, ethical approval was not required.

Measures

Global Quality Score

The GQS is a scale designed as an evaluation tool for online sources. The lowest score is 1 ("poor quality, poor flow of the site, most information missing, not at all useful for patients"), and the highest score is 5 ("excellent quality and excellent flow, very useful for patients"). Researchers use this scale to assess the flow, usability, and quality. A score of 4 or 5 is considered high quality, a score of 3 is considered moderate quality, and scores of 1 or 2 are considered low quality (29).

The Patient Education Materials Assessment Tool

The PEMAT was developed by Shoemaker et al. (37) in 2014 to evaluate the understandability and actionability of print and audiovisual patient education materials. PEMAT uses an inventory of both desirable and undesirable features of patient education materials to generate separate scores for comprehensibility and usability, ranging from 0 to 100. Each item on the scale is evaluated with a score of 0 (disagree) or 1 (agree), and some items have a third option, "no assessment", if applicable.

Table 1. Sample questions from each category

Topics	Sample questions
Symptoms	"What are some symptoms of autism that parents and caregivers can look for?"
Diagnosis	"How do health care providers diagnose autism?"
Treatment	"Are there treatments available for autism?"
General information	"How common is autism?"

PEMAT has two versions: PEMAT-P for print materials and PEMAT-A/V for audiovisual materials. In our study, PEMAT-P was used for evaluation. PEMAT-P includes 17 items for measuring understandability and 7 items for assessing actionability. Since the materials we evaluated and our study objectives do not focus on assessing any action, we planned to use only the 17 items related to understandability. The items in the scale are divided into six categories: content (e.g., "this material makes its purpose completely evident".), and word choice and style (e.g., "the material uses common, everyday language".), use of numbers (e.g., "numbers appearing in the material are clear and easy to understand."), organization (e.g., "the material presents information in a logical sequence".), layout and design (e.g., "the material uses visual cues to draw attention to key points".), and use of visual aids (e.g., "the material's visual aids have clear titles or captions".). The validity and reliability of the Turkish version of the assessment tool were established by Paylan Akkoç and Orgun (38) in 2020.

Statistical Analysis

Statistical analyses were implemented using SPSS version 28. The sum of the scores from the three researchers represented the total scores for each question. The mean total scores of the questions in general and each categorized topic were compared between the four Chatbots. The continuous variables were analyzed using one-one way ANOVA. Post-hoc comparisons were conducted using the Bonferroni-corrected t-tests. The findings of the variables were expressed as mean and standard deviation. Statistical significance was considered $p<0.05$.

RESULTS

Without categorizing by topic, there was a significant difference in word count between Gemini (197.47) and ChatGPT-3.5 (290.26) (p=0.009), as well as between Gemini (197.47) and ChatGPT-4 (306.74), (p=0.001) (Figure 1). Although no significant differences were found between the groups in terms of overall PEMAT understandability percentages, significant differences emerged when evaluating PEMAT subscale scores (Table 2). Specifically, content scores differed significantly between Microsoft Copilot (4.58) and ChatGPT-4 (5.68) (p=0.026.) Likewise, word choice and style scores showed a significant difference between Gemini (8.00) and ChatGPT-4 (6.21) (p=0.041). No significant differences were observed between the groups in the remaining PEMAT subscales. GQS differed significantly between Microsoft Copilot (9.11) and ChatGPT-4 (12.16) (p=0.006), as well as between Gemini (9.26) and ChatGPT-4 (12.16) (p=0.009), (Figure 2).

For the general information category, the only noteworthy difference observed was in word choice and style scores, with Gemini (9) outperforming ChatGPT-4 (6) (p=0.20). In the diagnosis category, GQS differed significantly between Microsoft Copilot (6.40) and ChatGPT-3.5 (11.20), (p=0.009), as well as between Microsoft Copilot (6.40) and ChatGPT-4 (12.80), (p<0.001).

There were no differences in the category of symptoms or treatment questions.

The analysis of the responses from chatbots (ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Microsoft Copilot) regarding common autism-related question themes is visualized in Figure 3. This stacked bar chart illustrates the contribution of each AI chatbot to the various identified themes, based on the frequency of relevant

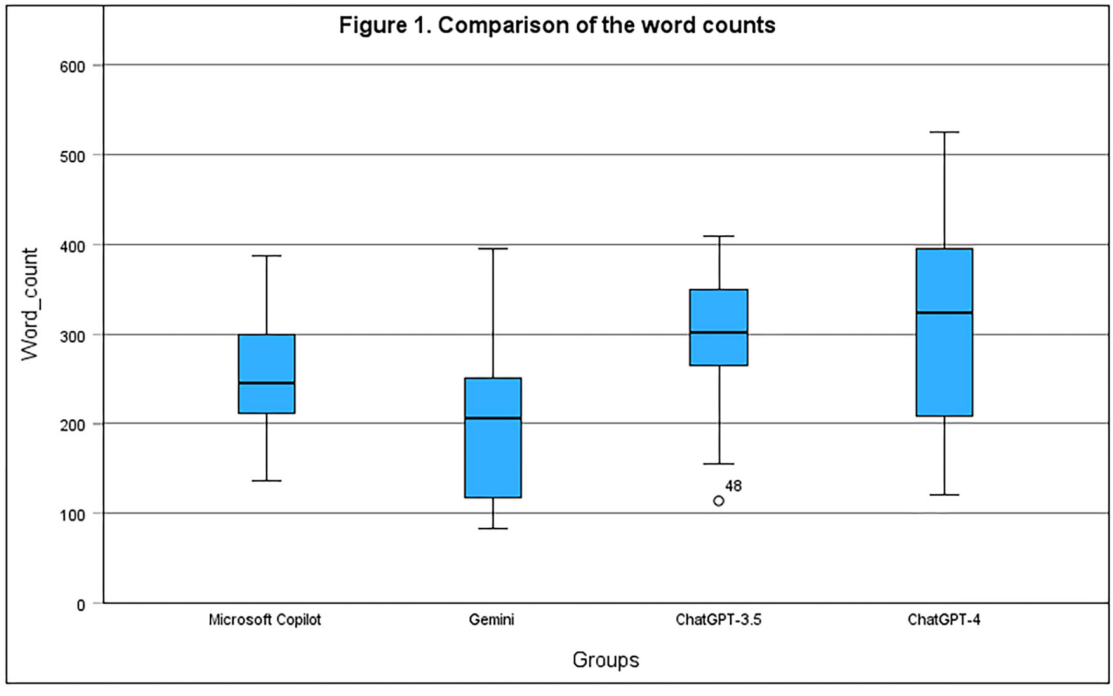

Figure 1. Comparison of word counts of answers to questions about ASD between chatbots.

ASD: Autism spectrum disorder

keywords. Each bar represents a theme, with different colors indicating the contributions from each chatbot. The total frequency of each theme is also labeled on the right of the bars.

Frequency distribution of content generated by four AI chatbots (GPT-3, GPT-4, Gemini, and Co-Pilot) across eight autism-related themes: early diagnosis, social challenges, communication difficulties, behavioral issues, intervention strategies, parental support, educational support, and therapeutic approaches. Values represent the number of chatbot responses assigned to each theme.

Table 3 below summarizes the themes identified from the responses to common autism-related questions. Each theme is associated with specific keywords, and the table indicates how frequently these keywords appear in the responses of each chatbot.

The sentiment analysis was conducted on the responses to common autism-related questions from ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Microsoft Copilot. The analysis aimed to determine the overall emotional tone of the responses by calculating average polarity (indicating positive or negative sentiment), and subjectivity (indicating the degree of personal opinion). Polarity scores range from -1 to 1, where -1 indicates a very negative sentiment, 0 indicates a neutral sentiment, and 1 indicates a very positive sentiment. The average polarity scores for all participants are slightly positive, indicating that the responses generally convey a positive sentiment towards the topics discussed. Subjectivity scores range from 0 to 1, where 0 indicates a fact-based response and 1 indicates a highly subjective or opinion-based response. The average subjectivity scores are moderately high, suggesting that the responses contain a mix of objective information and personal opinions or interpretations.

Table 2. Comparison of mean scores of PEMAT across groups using one-way ANOVA

Mean (SD)	Microsoft Copilot	Gemini	ChatGPT-3.5	ChatGPT-4	F	p-value	n_p^2
Understandability percentage	79.53 (10.46)	82.16 (9.50)	77.84 (9.30)	75.79 (9.72)	1.45	0.235	0.06
Content	4.58 (1.78)	5.21 (1.18)	5.53 (0.91)	5.68 (0.48)	3.25	0.026*	0.12
Word choice and style	6.63 (2.03)	8.00 (1.67)	6.58 (1.58)	6.21 (2.68)	2.83	0.044*	0.11
Use of numbers	4.42 (1.54)	4.26 (1.52)	3.63 (1.26)	4.26 (1.52)	1.09	0.361	0.04
Organization	8.74 (1.94)	7.89 (2.64)	7.95 (1.39)	7.68 (1.80)	1.02	0.390	0.04
Layout and design	2.84 (0.69)	2.53 (1.12)	2.21 (1.36)	2.05 (1.43)	1.65	0.185	0.06

*Post-hoc comparisons were conducted using the Bonferroni-corrected post-hoc t-tests. All error bars represent s.e.m. Significant results are bolded ($p<0.05$).
PEMAT: Patient education material assessment tool, SD: Standard deviation

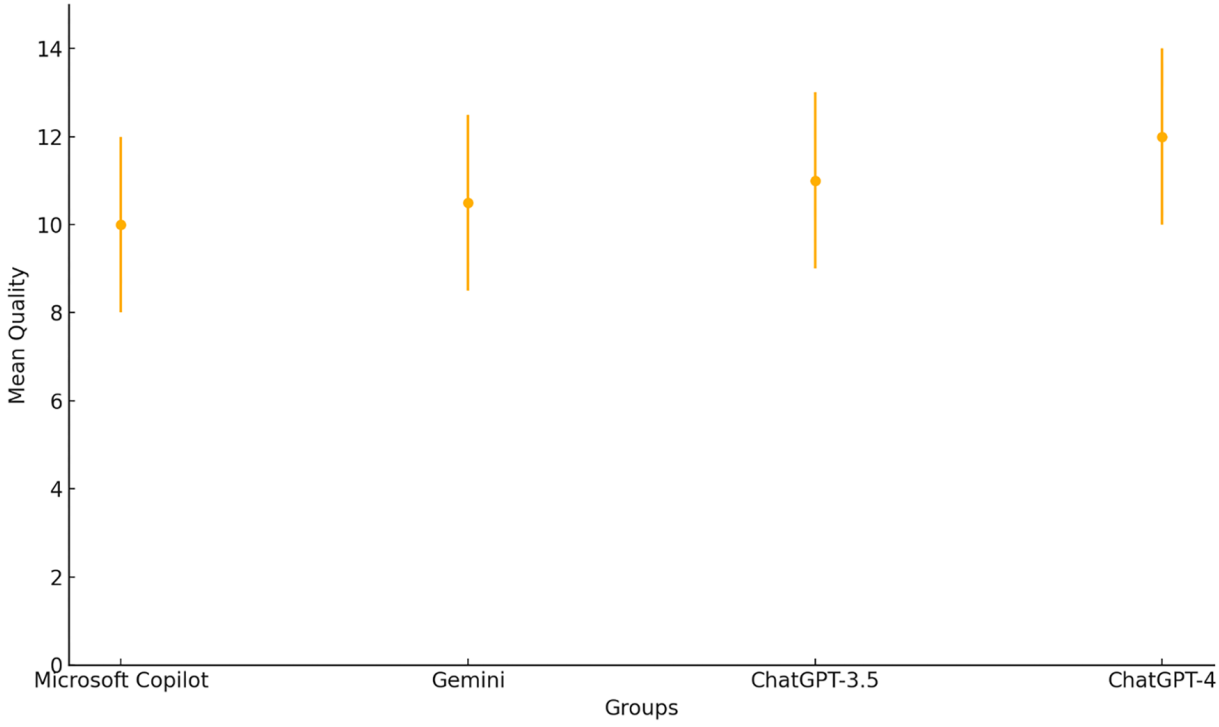

Figure 2. Comparison of Quality Scores of answers to questions about ASD between chatbots. Error bars represent 95% confidence intervals.
ASD: Autism spectrum disorder

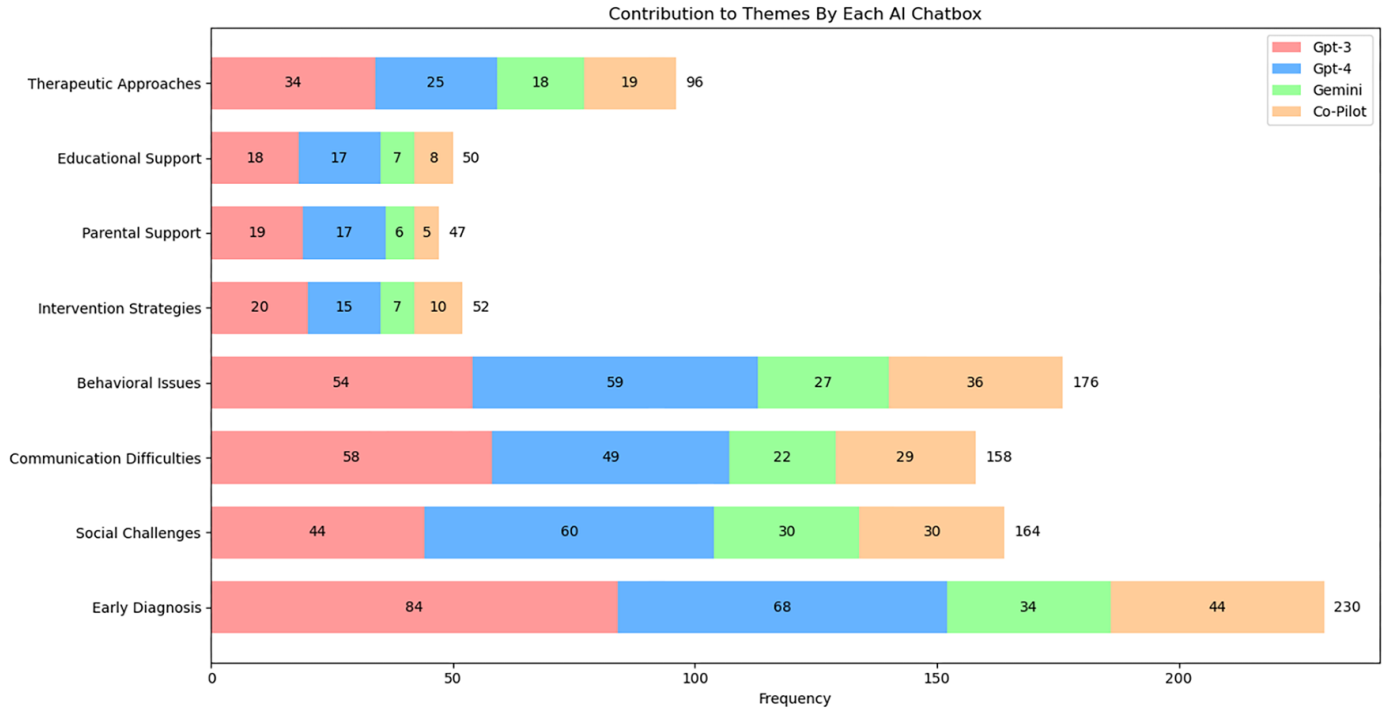

Figure 3. Themes and contribution to themes by each AI chatbot.
AI: Artificial intelligence

Table 3. Theme analysis by chatbots

Theme	Keywords	Frequency	ChatGPT-3	ChatGPT-4	Gemini	Copilot
Early diagnosis	Early, diagnose	230	84	68	34	44
Social challenges	Social	164	44	60	30	30
Communication difficulties	Communication	158	58	49	22	29
Behavioral issues	Behavior	176	54	59	27	36
Intervention strategies	Intervention, therapy	61	20	15	7	10
Parental support	Parent	47	19	17	6	5
Educational support	School, education	50	18	17	7	8
Therapeutic approaches	Therapy	96	34	25	18	19

Table 4. Polarity analysis of chatbots

AI chatbot	Polarity	Subjectivity
ChatGPT-3.5	0.092	0.442
ChatGPT-4.0	0.088	0.452
Gemini	0.115	0.436
Microsoft Copilot	0.088	0.446

AI: Artificial intelligence

DISCUSSION

This study analyzed the responses of several prominent chatbots, including ChatGPT-3.5, ChatGPT-4.0, Microsoft Copilot, and Google Gemini, to the most frequently asked questions regarding ASD. While the overall understandability scores, evaluated using the PEMAT tool, were comparable among chatbots, notable disparities were observed in the subscales of content, word choice, and style. While ChatGPT achieved a significantly higher score than Microsoft Copilot in the content subscale, Gemini also outperformed ChatGPT in word choice and style. ChatGPT had much better overall quality ratings compared to Microsoft Copilot and Gemini. Moreover, a thematic analysis of the AI-driven chatbot's written responses revealed that issues associated with "early diagnosis" were the most frequently emphasized. Sentiment analysis of responses from various chatbots consistently revealed a high degree of objectivity, with minimal polarity of emotions and a consistent neutral stance. The findings underscore the potential of AI systems to provide understandable and high-quality information, particularly regarding ASD, for individuals seeking such information. Nevertheless, this potential is not without its constraints. To the best of our knowledge, this study is the first to evaluate the responses generated by widely used chatbots to frequently asked questions about autism using both quantitative and qualitative methods. Furthermore, the data for this study was collected on world autism awareness day aims to raise awareness about autism from a different perspective.

In our study, the understandability scores assessed through 17 items on the PEMAT revealed that all chat bots demonstrated similar scores, generally producing comprehensible responses. These results are consistent with the study by McFayden et al. (36) on the responses of ChatGPT-4.0 to autism inquiries, which found similar understandability. Notwithstanding similar levels of understandability, ChatGPT-4.0 outperformed Microsoft Copilot in terms of the content subscale, while Google Gemini achieved higher scores than ChatGPT-4.0 in terms of word choice and style, subscale. This indicates that although overall understandability is similar, there are notable differences in the depth of content and linguistic precision each platform offers. Thus, our study, by evaluating not only ChatGPT but also other commonly used LLMs, enables a more comprehensive understanding of the relative strengths and weaknesses of these models in the dissemination of ASD-related health information. This study adds to the growing body of evidence indicating that in poorly resourced settings, AI-driven tools could have potential application for public health education, especially in scenarios where access to professional healthcare is compromised.

Among the evaluated AI-driven chatbots, ChatGPT-4.0 consistently stood out by providing responses to common ASD-related questions, and achieved the highest average overall quality score, with statistically significant differences compared to Microsoft Copilot and Google Gemini. These findings align with previous studies, which also emphasized GPT-4.0's superiority in radiological decision-making and its responses to myopia-related queries (34,39). Other studies have further highlighted GPT-4.0's reliability and depth in addressing complex medical conditions, supporting the potential of this technology in disseminating medical information (40).

The current variability in performance within LLMs, such as ChatGPT-4, Google Gemini, and Microsoft Copilot, is primarily due to architectural differences and the datasets these models have been exposed to. A few other reasons contribute to better performance in ChatGPT-4, especially in giving high-quality and more detailed responses regarding ASD. The difference between GPT-4 and the older versions, such as GPT-3.5, is that it has many more parameters and makes use of much more advanced transformer architectures (41). This means it will be able to learn even more complicated patterns in language and then reproduce them, thus giving more subtle and contextually appropriate answers, especially in medical contexts. Apart from that, GPT-4 has undergone extensive fine-tuning, especially by Reinforcement Learning from Human Feedback, which enhances its potential for responses in line with human-like values of empathy; fine-tuning is useful, especially in sensitive topics such as autism, where it is imperative to consider tone and factual accuracy. Other models such as Google Gemini and Microsoft Copilot, however, though very powerful with respect to general tasks, have not been as thoroughly fine-tuned in domain-specific contexts like healthcare. While Google Gemini does a good job in terms of choosing words and style, its interest seems more in linguistic refinement than the actual content accuracy observed with GPT-4 (42). Domain-specific knowledge integration likely varied during training. Microsoft's Copilot is also not well-suited for medical guidance, for which it is not optimized; it is more biased toward tasks and code-driven applications. These architectural differences affect information quality, especially in specialized domains like ASD. Similar significant values associated with these findings indicate that domain-specific training and fine-tuning of LLMs for practical applications is mandatory, especially in healthcare, where accuracy, empathy, and contextual relevance of responses remain critical (41).

Considering the rising incidence of ASD worldwide (4), the gross inadequacy of mental health professionals, especially in low- and middle-income countries (5), as well as ongoing stigma against neurodevelopmental disorders like ASD (9,11,43), our results indicate that AI-chatbots could be pivotal in addressing these needs. Specifically, ChatGPT-4.0's ability to provide detailed and comprehensible information can be highly beneficial in closing the knowledge gap for families and healthcare providers, especially in regions with limited access to mental health services. However, it is important to note from our findings, that ChatGPT-3.5 also demonstrated comparable performance to ChatGPT-4.0 in responding to common ASD-related questions. For lower-middle-income countries, where access to more advanced models like ChatGPT-4.0 could be heavily restricted by monetary barriers, ChatGPT-3.5 might even be considered a relatively cheap alternative. This provides a reminder to evaluate both the cost and performance factors when considering AI-driven technologies to achieve healthcare access parity across socioeconomic contexts. Future studies should investigate the adaptation of these tools in clinical practice, within regions where health care access is challenging, and understanding outcomes on a real-world basis, including patient-reported benefits, while also exploring increasing information dissemination.

In the results by category, both ChatGPT-3.5 and ChatGPT-4.0 had statistically higher overall quality scores for the "diagnosis" compared to those of other chatbots. This finding is particularly important because families often seek information at critical moments when their child has either been diagnosed with ASD or when they suspect ASD (44). In these circumstances, access to reliable and accurate

diagnostic information is essential. Early diagnosis significantly minimizes the delay in intervention, thereby enhancing the long-term developmental outcomes for children with ASD (45). Thus, inaccurate data and ambiguity in sources can lead to considerable delays in the diagnostic process (46). In the context of ASD where early diagnosis and intervention are key to better outcomes, having accurate and understandable information widely available is incredibly important. As a consequence, if patients or caregivers utilize poorly structured or deceptive advice based on information retrieved from AI-driven chatbots or other online sources, patients and caregivers may be left confused, which might impede them from reaching crucial medical consultation. This highlights the need for AI systems not only to provide accurate, but also understandable medical content that directs users to proper clinical care. Thus, it is advisable that chatbots maintain updates of their diagnostic data and use technologies that provide readability to this available information so that they provide accurate information and make their users interact accordingly. Future research should explore how these AI-powered tools can be optimized to provide more specific and context-based information for both families and professionals.

A thematic analysis of the responses generated by AI-powered chatbots identified "early diagnosis" as the most frequently emphasized keyword, underscoring the critical importance of early intervention in ASD. The existing literature extensively documents that early diagnosis, by enabling timely and effective interventions, can significantly improve developmental outcomes (47). As such, it has become a fundamental theme for families seeking information regarding ASD (44). The prominence given to early diagnosis by ChatGPT-3.5 and ChatGPT-4.0, compared with other chatbots, reflects a notable strength of these models. It is recommended that other chatbots, particularly Microsoft Copilot and Google Gemini, prioritize integrating early diagnosis into their content to enhance the effectiveness of public health messaging.

In addition to early diagnosis, other key themes identified included social challenges and communication difficulties. ChatGPT-4.0 placed greater emphasis on social challenges, whereas ChatGPT-3.5 had a greater focus on communication difficulties, reflecting the necessity of targeted interventions in these core areas of ASD. These findings suggest that AI-powered chatbots not only provide general information but can also be optimized to offer more specific and contextual guidance regarding the distinct challenges faced by individuals with ASD and their families.

Behavioral issues also emerged as a critical theme, highlighted by the contributions of both ChatGPT-4.0 and ChatGPT-3.5. This emphasizes the importance of behavioral interventions in the effective management of behaviors associated with autism. The ability of these chatbots to recognize the variability in ASD symptoms is particularly significant, aligning with the current clinical understanding that no two individuals with autism present identical behavioral profiles (48). This recognition underscores the necessity for personalized therapeutic approaches in the diagnosis and treatment of individuals with ASD.

Furthermore, intervention strategies and therapeutic approaches were identified as salient themes, with ChatGPT-3.5 demonstrating superior performance in discussions of various therapeutic interventions. The emphasis on therapeutic approaches is particularly relevant, as individualized therapies-whether applied behavior analysis, medication, cognitive therapy, or sensory-based-are crucial for addressing the specific needs of individuals with ASD (49,50). Parental and educational support, extensively recognized in the literature as essential components in the effective management of autism, further enhances the value of the information generated by these chatbots.

Sentiment analysis indicated that nearly all chatbots employed a neutral, objective emotional tone when their polarity and subjectivity were measured. The results are both predictable and acceptable, given that the study evaluated chatbot responses related to health conditions. This serves as a crucial reminder to provide health education based on objective and factual information, particularly in discussions regarding clinical conditions such as ASD (51). This resource will assist families navigating the complex and emotionally burdening process of ASD by providing accurate and impartial information to empower them. The relatively high subjectivity scores may stem from the nature of LLMs' training data, which largely consists of human-authored, interpretive texts rather than strict clinical guidelines. Additionally, the responses of the models are not presented in a formal academic format but instead adopt an explanatory style for general audiences, which inherently incorporates more interpretive language.

There are several limitations that should be considered when interpreting the findings of our study. First, LLMs are dynamic systems that continually update their data and adjust their responses based on user interactions. As a result, the responses analyzed in our study may differ from those generated before or after when the questions were asked. This makes it harder to maintain chatbots' output consistently over time, especially when additional information is included in their databases. Future work can study what impact these updates have on the quality and accuracy of answers, especially for urgent health-related questions. Second, we attempted to capture a broad sample by aggregating questions across platforms, but the questions included may not cover all possible concerns that families have in practice. The same applies in the scenario of ASD, which is also context-specific because there may be different concerns depending on individual cases and family dynamics. Investigators should aim to enhance the generalizability of their data by including additional sources of feedback, such as caregivers, as well as clinicians, in future studies. Lastly, yet the questions were posed in English, therefore making it difficult to generalize our findings to non-English-speaking populations. The language in which a chatbot response is formulated can have a major impact on its clarity, and the effectiveness of these mechanisms requires further research depending on the cultural and linguistic contexts. This is particularly relevant in areas with poor health services and the promotion of information on healthcare, which could be a key role for chatbots. This limitation must be mitigated to understand the potential generalizability of LLMs as global health information aids.

In conclusion, our study presented an evaluation of responses to frequently asked questions about ASD using four of the most used AI-powered chatbots. The responses were rated similarly on overall understandability across all chatbots but varied on two sub-dimensions: content and word choice. When assessed in terms

of overall response quality, ChatGPT-4.0 demonstrated superior performance compared to Microsoft Copilot and Google Gemini. As AI increasingly influences the dissemination of health content, it becomes essential that the information provided by these platforms is both accurate and precise. For effective delivery of health tools to the public, it is essential that chatbots offer real-time, scientifically grounded health information. The next step involves evaluating the enduring efficacy of AI-driven chatbots and investigating whether modifications in machine learning models result in enhanced information quality. It is necessary to evaluate the usability of these tools across various languages and cultures. This is essential for understanding the potential impact of AI on global health challenges and for addressing inequities in access to health information.

Ethics

Ethics Committee Approval: Since our study was not a study involving humans and animals, ethics committee approval was not required.

Informed Consent: Patient consent was not required.

Footnotes

Authorship Contributions

Surgical and Medical Practices: G.D., C.D.H., Concept: D.B., A.Ö., Design: D.B., A.Ö., Data Collection or Processing: G.D., M.S., C.D.H., Analysis or Interpretation: M.S., D.B., A.Ö., Literature Search: G.D., C.D.H., Writing: G.D., C.D.H.

Conflict of Interest: Ahmet Özaslan, MD, serves as Section Editor in Internal Medicine for the Gazi Medical Journal. He had no involvement in the peer-review of this article and had no access to information regarding its review process. Other authors have nothing to disclose.

Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. VA: American Psychiatric Association. 2013. Available from: https://doi.org/10.1176/appi.books.9780890425596
2. Hall CM, Culler ED, Frank-Webb A. Online dissemination of resources and services for parents of children with autism spectrum disorders (ASDs): a systematic review of evidence. Review Journal of Autism and Developmental Disorders. 2016; 3: 273-85.
3. Rizzo A, Sorrenti L, Commendatore M, Mautone A, Caparello C, Maggio MG, et al. Caregivers of children with autism spectrum disorders: the role of guilt sensitivity and support. J Clin Med. 2024; 13: 4249.
4. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring Network, 11 Sites, United States, 2020. MMWR Surveill Summ. 2023; 72: 1-14.
5. World Health Organization. World mental health report: transforming mental health for all. Geneva: World Health Organization; 2022. Available from: https://www.who.int/publications/i/item/9789240063600
6. Srinivasan S, Ekbladh A, Freedman B, Bhat A. Needs assessment in unmet healthcare and family support services: a survey of caregivers of children and youth with autism spectrum disorder in Delaware. Autism Res. 2021; 14: 1736-58.
7. Courcy I, des Rivières-Pigeon C. "We're responsible for the diagnosis and for finding help". The help-seeking trajectories of families of children on the autism spectrum. Sociol Health Illn. 2021; 43: 40-57.
8. Freund PE, McGuire MB, Podhurst LS. Health, illness, and the social body: a critical sociology. Prentice Hall. 2003.
9. Broady TR, Stoyles GJ, Morse C. Understanding carers' lived experience of stigma: the voice of families with a child on the autism spectrum. Health Soc Care Community. 2017; 25: 224-33.
10. Courcy I, des Rivières C. "From cause to cure": a qualitative study on contemporary forms of mother blaming experienced by mothers of young children with autism spectrum disorder. Journal of Family Social Work. 2017; 20: 233-50.
11. Papadopoulos C, Lodder A, Constantinou G, Randhawa G. Systematic review of the relationship between autism stigma and informal caregiver mental health. J Autism Dev Disord. 2019; 49: 1665-85.
12. Lacruz-Pérez I, Sanz-Cervera P, Pastor-Cerezuela G, Gómez-Marí I, Tárraga-Mínguez R. Is it possible to educate, intervene or "cure" autism spectrum disorder? A content analysis of YouTube videos. Int. J. Environ. Res. Public Health. 2021; 18: 2350.
13. Bellon-Harn ML, Manchaiah V, Morris LR. A cross-sectional descriptive analysis of portrayal of autism spectrum disorders in YouTube videos: a short report. Autism. 2020; 24: 263-8.
14. Ozgor F, Caglar U, Halis A, Cakir H, Aksu UC, Ayranci A, et al. Urological cancers and ChatGPT: assessing the quality of information and possible risks for patients. Clin Genitourin Cancer. 2024; 22: 454-7.
15. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak. 2021; 21: 125.
16. Bhardwaz S, Kumar J. An extensive comparative analysis of chatbot technologies-ChatGPT, Google BARD and Microsoft Bing. 2023 2nd international conference on applied artificial intelligence and computing (ICAAIC). 2023.
17. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023; 11: 1166120.
18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023; 388: 1233-9.
19. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023; 11: 887.
20. Biswas SS. Role of Chat GPT in public health. Ann Biomed Eng. 2023; 51: 868-9.
21. Reavley NJ, Logan N, Morgan A, Ross A, Jorm AF. Research letter: could ChatGPT and Bard provide helpful responses to a person seeking advice on how to support someone with a mental health problem? Aust N Z J Psychiatry. 2024; 58: 373-5.
22. Shahsavar Y, Choudhury A. User intentions to use ChatGPT for self-diagnosis and health-related purposes: cross-sectional survey study. JMIR Hum Factors. 2023; 10: e47564.
23. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (turing) test: survey study. JMIR Med Educ. 2023; 9: e46939.
24. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ. 2023; 9: e51243.

25. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry. 2023; 22: 479-80.
26. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023; 183: 589-96.
27. Ilicki J. A Framework for critically assessing ChatGPT and other large language artificial intelligence model applications in health care. Mayo Clin Proc Digit Health. 2023; 1: 185-8.
28. Carlbring P, Hadjistavropoulos H, Kleiboer A, Andersson G. A new era in internet interventions: the advent of Chat-GPT and AI-assisted therapist guidance. Internet Interv. 2023; 32: 100621.
29. Botha M, Hanlon J, Williams GL. Does language matter? Identity-first versus person-first language use in autism research: a response to vivanti. J Autism Dev Disord. 2023; 53: 870-8.
30. Vivanti G. Ask the Editor: What is the most appropriate way to talk about individuals with a diagnosis of autism? J Autism Dev Disord. 2020; 50: 691-3.
31. Gernsbacher MA. Editorial perspective: the use of person-first language in scholarly writing may accentuate stigma. J Child Psychol Psychiatry. 2017; 58: 859-61.
32. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023; 29: 721-32.
33. 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.
34. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, 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: 104770.
35. Aguirre A, Hilsabeck R, Smith T, Xie B, He D, Wang Z, et al. Assessing the quality of ChatGPT responses to dementia caregivers' questions: qualitative analysis. JMIR Aging. 2024; 7: e53019.
36. McFayden TC, Bristol S, Putnam O, Harrop C. ChatGPT: artificial intelligence as a potential tool for parents seeking information about autism. Cyberpsychol Behav Soc Netw. 2024; 27: 135-48.
37. Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns. 2014; 96: 395-403.
38. Paylan Akkoç C, Orgun F. Psychometric testing of the turkish version of the patient education materials assessment tool. Florence Nightingale J Nurs. 2023; 31: 180-7.
39. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol. 2023; 20: 990-7.
40. Deng L, Wang T, Yangzhang, Zhai Z, Tao W, Li J, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg. 2024; 110: 1941-50.
41. Liu CL, Ho CT, Wu TC. Custom GPTs enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? A study on the emergency medicine specialist examination. Healthcare (Basel). 2024; 12: 1726.
42. Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large language models for intraoperative decision support in plastic surgery: a comparison between ChatGPT-4 and gemini. Medicina (Kaunas). 2024; 60: 957.
43. Özaslan A, Yıldırım M. Internalized stigma and self-esteem of mothers of children diagnosed with attention deficit hyperactivity disorder. Children's Health Care. 2021; 50: 312-24.
44. Grant N, Rodger S, Hoffmann T. Intervention decision-making processes and information preferences of parents of children with autism spectrum disorders. Child Care Health Dev. 2016; 42: 125-34.
45. Hadders-Algra M. Early Diagnostics and early intervention in neurodevelopmental disorders-age-dependent challenges and opportunities. J Clin Med. 2021; 10: 861.
46. Gabis LV, Attia OL, Goldman M, Barak N, Tefera P, Shefer S, et al. The myth of vaccination and autism spectrum. Eur J Paediatr Neurol. 2022; 36: 151-8.
47. Sapiets SJ, Totsika V, Hastings RP. Factors influencing access to early intervention for families of children with developmental disabilities: A narrative review. J Appl Res Intellect Disabil. 2021; 34: 695-711.
48. Brigido E, Rodrigues A, Santos S. Autism spectrum disorder behavioral profiles: a cluster analysis exploration. International Journal of Disability, Development and Education. 2023; 70: 515-29.
49. Frye RE. A Personalized multidisciplinary approach to evaluating and treating autism spectrum disorder. J Pers Med. 2022; 12: 464.
50. McMahon CM, McClain MB, Wells S, Thompson S, Shahidullah JD. Autism knowledge assessments: a closer examination of validity by autism experts. J Autism Dev Disord. 2025; 55: 1629-47.
51. Bottema-Beutel K, Kapp SK, Lester JN, Sasson NJ, Hand BN. Avoiding ableist language: suggestions for autism researchers. Autism Adulthood. 2021; 3: 18-29.