DOI: 10.4274/jcrpe.galenos.2025.2025-6-14

Research Article

Responses of Different Artificial Intelligence Systems to Questions Related with Short Stature as Assessed by Pediatric Endocrinologists

Kaşali K et al. Responses of Different Artificial Intelligence Systems to Questions Related with Short Stature as Assessed by Pediatric Endocrinologists

Kamber Kaşali[1], Özgür Fırat Özpolat[2], Merve Ülkü[3], Ayşe Sena Dönmez[4], Serap Kılıç Kaya[4], Esra Dişçi[4], Serkan Bilge Koca[5], Ufuk Özkaya[4], Hüseyin Demirbilek[6], Atilla Çayır[7]
[1]Department of Biostatistics, Atatürk University Faculty of Medicine, Erzurum, Türkiye
[2]Data Management Office, Atatürk University, Erzurum, Türkiye
[3]Erzurum City Hospital, Clinic of Pediatric Endocrinology, Erzurum, Türkiye
[4]Erzurum City Hospital, Clinic of Pediatrics, Erzurum, Türkiye
[5]Department of Pediatrics, Division of Pediatric Endocrinology, Kayseri City Training and Research Hospital, Kayseri, Türkiye
[6]Hacettepe University Faculty of Medicine, Department of Pediatric Endocrinology, Ankara, Türkiye
[7]Atatürk University Faculty of Medicine, Department of Pediatric Endocrinology, Erzurum, Türkiye

What is already known about this?

Artificial intelligence (AI) is increasingly used in medical decision-making, including in paediatric endocrinology. AI models can help diagnose short stature by analysing growth patterns and related factors, but not much is known about their accuracy and reliability.

What does this study add?

This study evaluates AI-generated answers on short stature by comparing them with expert opinions. It highlights the strengths and limitations of AI in clinical decision-making and identifies areas where AI is or is not in line with expert recommendations.

Abstract

Objective: Artificial intelligence (AI) is increasingly utilized in medicine, including pediatric endocrinology. AI models have the potential to support clinical decision-making, patient education, and guidance. However, their accuracy, reliability, and effectiveness in providing medical information and recommendations remain unclear. This study aims to evaluate and compare the performance of four AI models—ChatGPT, Bard, Microsoft Copilot, and Pi—in answering frequently asked questions related to pediatric endocrinology.

Methods: Nine questions commonly asked by parents regarding short stature in paediatric endocrinology have been selected based on literature reviews and expert opinions. These questions were posed to four AI models in both Turkish and English. The AI-generated responses were evaluated by 10 pediatric endocrinologists using a 12-item Likert-scale questionnaire assessing medical accuracy, completeness, guidance, and informativeness. Statistical analyses, including Kruskal-Wallis and post-hoc tests, were conducted to determine significant differences between AI models.

Results: Bard outperformed other models in guidance and recommendation categories, excelling in directing users to medical consultation. Microsoft Copilot demonstrated strong medical accuracy but lacked guidance capacity. ChatGPT showed consistent performance in knowledge dissemination, making it effective for patient education. Pi scored the lowest in guidance and recommendations, indicating limited applicability in clinical settings. Significant differences were observed among AI models ($p < 0.05$), particularly in completeness and guidance-related categories.

Conclusion: The study highlights the varying strengths and weaknesses of AI models in pediatric endocrinology. While Bard is effective in guidance, Microsoft Copilot excels in accuracy, and ChatGPT is informative. Future AI improvements should focus on balancing accuracy and guidance to enhance clinical decision-support and patient education. Tailored AI applications may optimize AI's role in specialized medical fields.

Keywords: Pediatric Endocrinology, Artificial Intelligence (AI), Clinical Decision Support, Medical Informatics

Asst. Prof. Kamber Kaşali, Department of Biostatistics, Atatürk University Faculty of Medicine, Erzurum, Türkiye
kmbrkasali76@gmail.com
https://orcid.org/0000-0002-2851-5263

11.06.2025
14.08.2025

Epub: 02.09.2025

Introduction

Artificial intelligence (AI) has been rapidly expanding its applications in the field of medicine, including pediatric endocrinology. The complexity of clinical problems and the rapidly evolving need for information in pediatric endocrinology further enhance the potential of AI in this domain. This study evaluates the responses provided by AI systems to frequently asked questions in pediatric endocrinology. The literature demonstrates the applicability of AI in various areas, including growth disorders, obesity, diabetes management, and hormonal imbalances (1,2,3).

The integration of AI into pediatric endocrinology has become particularly prominent in diabetes management. Winkelman and friends reported in 2023 that AI has been successfully utilized in optimizing insulin dosing and predicting hypoglycemia risk (4). Additionally, Zhang et al. found that AI-assisted bone age analyses improve diagnostic accuracy in cases of growth hormone deficiency (3).

AI has also shown significant contributions to the early diagnosis of thyroid diseases. Otjen et al. highlighted the high success rate of AI in the automated analysis of thyroid ultrasound images (5). Furthermore, AI models used in obesity and insulin resistance management facilitate personalized treatment approaches (1,2).

UNCORRECTED PROOF

Regarding growth disorders, the accuracy of AI in bone age measurement and its impact on accelerated diagnostic processes are particularly noteworthy. Waikel et al. emphasized that AI serves as an effective educational tool in recognizing genetic syndromes (6).

This study aims to analyze the accuracy of AI-generated responses to the aforementioned questions and their evaluation by expert pediatric endocrinologists. The integration of AI into clinical practice has the potential to reduce healthcare professionals' workload while playing a crucial complementary role in patient care and clinical decision-making. However, challenges such as data security, ethical concerns, and algorithmic accuracy remain critical issues that need to be addressed.

Materials and Methods

This study aimed to evaluate the responses provided by artificial intelligence (AI) systems to questions frequently asked by the parents of pediatric endocrinology patients, based on expert assessment.

In our study, a literature review and expert opinions were utilized to identify the nine most frequently asked questions, which were then posed to AI models. Subsequently, the AI-generated responses were evaluated by 10 pediatric endocrinologists. A 12-item questionnaire was developed to assess these responses, and the endocrinologists were asked to complete it.

Participants

The study included 10 pediatric endocrinologists specializing in pediatric endocrinology. The participants were selected randomly (using a simple random sampling method) from experts who had at least five years of experience in pediatric endocrinology and were actively engaged in clinical practice. These paediatricians are not among the authors.

Question Development

To determine the most commonly asked questions by the parents of pediatric endocrinology patients, a literature review was conducted, and expert opinions were sought. As a result, a total of nine questions were formulated. Each AI model was queried separately in both Turkish and English. The selected questions were as follows:

1. What is short stature?
2. What are the causes of short stature?
3. How is growth velocity assessed in short stature?
4. How is bone age determined in cases of short stature?
5. What should be considered in the differential diagnosis of short stature?
6. Which laboratory parameters should be evaluated in cases of short stature?
7. What medications are used in the treatment of short stature?
8. How frequently should short stature be monitored?
9. What are the potential side effects of growth hormone therapy?

Artificial Intelligence Models

The questions were posed to four different AI models: ChatGPT, Bard, Microsoft Copilot, and Pi. Each AI model was queried separately in both Turkish and English, and the responses were recorded for further analysis.

Evaluation Process

The responses obtained from AI systems were evaluated by 10 pediatric endocrinologists. A 12-item Likert-type questionnaire was used for the assessment. For each AI-generated response, experts rated the following survey questions on a scale from 1 to 5:

1. Was a proper definition provided?
2. Was all necessary information included?
3. Was any essential information missing?
4. Was excessive information provided?
5. Was any irrelevant information included?
6. Was the medical information accurate?
7. Were recommendations given?
8. Was patient guidance provided?
9. Was a recommendation to consult a physician included?
10. Was the response sufficient for the patient?
11. Did the response aim to inform the reader?
12. Did the response aim to reassure the reader?

Statistical Analysis

The data were presented as mean, standard deviation, median, minimum, and maximum values. The obtained data were analyzed using SPSS 20.0 software. The Kruskal-Wallis test was used to determine the significance of differences between the responses to the questions. In cases where significant differences were observed, post-hoc tests were conducted using the Kruskal-Wallis one-way ANOVA (k samples) test. A significance level of *p*<0.05 was considered statistically significant.

Results

A total of 10 pediatric endocrinologists specializing in pediatric endocrinology participated in the study.

Table 1 presents the evaluation results of responses provided by four AI models (ChatGPT, Bard, Microsoft Copilot (M.C.), and Pi) to nine pediatric endocrinology related questions, as assessed by experts using a 12-item Likert type questionnaire. The expert evaluation results for each question posed to AI models are summarized as follows.

Evaluation of AI Models in Answering "What is Short Stature?"

Bard received the highest score for definition accuracy (5 ± 1), though the difference among models was not statistically significant (p = 0.139). In the "Missing Information Provided" category, ChatGPT had a higher tendency for incomplete responses compared to Bard (**p = 0.027**). Bard and MC performed best in "Recommendations Provided" and "Patient Guidance Provided" while Pi scored the lowest (**p < 0.001**). Bard and MC also excelled in "Response Aims to Inform the Reader" with MC significantly outperforming ChatGPT (**p = 0.007**). In "Response Aims to Reassure the Reader" Bard led, and Pi ranked lowest, with a significant difference between Bard and MC (**p < 0.001**) (Table 1).

Evaluation of AI Models in Answering "What Are the Causes of Short Stature?"

ChatGPT scored highest in the "All Necessary Information Provided" category (4 ± 1), significantly outperforming MC and Pi (**p = 0.001**). In "Missing Information Provided" Bard had the lowest score (2 ± 1), with Pi and MC scoring higher (**p = 0.011**). Bard and ChatGPT performed better in avoiding irrelevant information compared to MC (**p = 0.011**). Bard excelled in "Recommendations Provided" (4 ± 1) and "Patient Guidance Provided" (5 ± 0), while Pi scored the lowest (**p < 0.001**). In "Recommendation to Consult a Physician" Bard and MC led, whereas Pi performed significantly worse (**p < 0.001**). ChatGPT, Bard, and MC scored highest in "Response Aims to Inform the Reader", with Pi lagging behind (**p = 0.041**) (Table 1).

UNCORRECTED PROOF

Evaluation of AI Models in Answering “How Is Growth Rate Evaluated in Short Stature?”

In the “Definition Provided” category, ChatGPT scored higher (4 ± 1) than Pi (3 ± 1), demonstrating superior definition clarity (**p = 0.038**). In “Missing Information Provided” Bard (2 ± 1) performed better than MC (4 ± 1), highlighting Bard’s ability to provide more complete responses (**p = 0.038**). Bard excelled in “Recommendations Provided” (4 ± 0), while MC scored the lowest (2 ± 1), with a significant difference (**p < 0.001**). In “Patient Guidance Provided” Bard (4 ± 1) outperformed MC (2 ± 1) (**p = 0.007**). Similarly, in “Recommendation to Consult a Physician” Bard (4 ± 1) led, while MC (2 ± 1) performed poorly (**p = 0.003**). Lastly, in “Response Was Sufficient for the Patient” MC had a significantly lower score than other models (**p = 0.038**), indicating its weaker performance in providing satisfactory responses (Table 1).

Evaluation of AI Models in Answering “How Is Bone Age Determined in Short Stature?”

In the “Definition Provided” category, no significant difference was found among the models (p = 0.423), with ChatGPT scoring highest (4 ± 1). In “Recommendations Provided” Bard (4 ± 1) significantly outperformed Pi and MC (2 ± 1) (**p < 0.001**), confirming its superiority in offering guidance. In “Patient Guidance Provided” Bard (4 ± 0) excelled, significantly outperforming all other models (**p < 0.001**). Similarly, in “Recommendation to Consult a Physician” Bard (5 ± 1) led, while Pi (2 ± 0) and MC (2 ± 1) performed the worst (**p < 0.001**). Finally, in “Response Was Sufficient for the Patient” Bard (4 ± 1) was the most effective, while MC (3 ± 1) scored significantly lower (**p = 0.042**), indicating Bard’s stronger ability to meet users' informational needs (Table 1).

Evaluation of AI Models in Answering “What Should Be Considered in Differential Diagnosis in Short Stature?”

In the “Definition Provided” category, ChatGPT scored highest (4 ± 1), significantly outperforming Bard (3 ± 1) (**p = 0.040**). In “Patient Guidance Provided” Bard and MC (4 ± 0) excelled, while Pi and ChatGPT (3 ± 1) performed lower, with significant differences between Pi-Bard and ChatGPT-Bard (**p = 0.005**). For “Recommendation to Consult a Physician” Bard and MC (5 ± 1) were the most effective, while Pi (3 ± 1) performed the weakest, with significant differences between Pi-MC and Pi-Bard (**p = 0.001**). In “Response Aims to Inform the Reader” ChatGPT, Bard, and MC (4 ± 1) performed well, whereas Pi (3 ± 1) lagged, showing a significant difference from ChatGPT (**p = 0.004**) (Table 1).

Table 2 presents the evaluation results of responses provided by four AI models (ChatGPT, Bard, MC, and Pi) to nine pediatric endocrinology-related questions, as assessed by experts using a 12-item Likert-type questionnaire. The expert evaluation results for each question posed to AI models are summarized as follows.

Evaluation of AI Models in Answering “What Should Be Considered in Laboratory Parameters in Short Stature?”

In the “Definition Provided” and “All Necessary Information Provided” categories, all models received similar scores, with no significant differences (p = 0.595 and p = 0.446, respectively). Although ChatGPT scored highest, the variations were not statistically significant. In “Patient Guidance Provided” Bard (4 ± 1) outperformed MC (2 ± 1) and Pi (3 ± 1), with a significant difference between MC and Bard (**p = 0.030**), indicating Bard’s stronger guidance ability. Similarly, in “Recommendation to Consult a Physician” Bard (3 ± 1) and ChatGPT (3 ± 1) scored higher than MC (2 ± 1), with Bard significantly outperforming MC (**p = 0.014**). For “Response Was Sufficient for the Patient” ChatGPT (4 ± 1) led, while MC (2 ± 1) and Bard (3 ± 1) scored lower, with a significant difference between MC and ChatGPT (**p = 0.018**). Lastly, in “Response Aims to Inform the Reader” ChatGPT (4 ± 1) outperformed MC (3 ± 1), with a significant difference (**p = 0.033**), confirming ChatGPT’s superior capacity for providing informative responses (Table 2).

Evaluation of AI Models in Answering “Which Drugs Are Used in the Treatment of Short Stature?”

In the “All Necessary Information Provided” category, MC scored the lowest (2 ± 1), significantly underperforming compared to Bard and ChatGPT (4 ± 1) (**p < 0.001**). This suggests MC was less effective in providing comprehensive information. In “Missing Information Provided” MC (4 ± 1) had the highest score, indicating a greater tendency to provide incomplete information. ChatGPT (2 ± 1) and Pi (2 ± 1) scored lower, with MC significantly differing from these models (**p = 0.002**). In “Medically Accurate Information Provided” MC (3 ± 1) slightly outperformed ChatGPT, with a statistically significant difference (**p = 0.036**), highlighting MC’s relative strength in medical accuracy. For “Recommendations Provided” Bard (4 ± 1) scored highest, with a significant difference from MC (3 ± 1) and Pi (3 ± 1) (**p = 0.027**), confirming Bard’s superiority in offering guidance. In “Response Was Sufficient for the Patient” ChatGPT (4 ± 1) led, while MC and Pi (2 ± 1) scored lower. MC performed significantly worse than ChatGPT (**p = 0.002**), demonstrating ChatGPT’s stronger ability to meet users' informational needs. Lastly, in “Response Aims to Inform the Reader” ChatGPT (4 ± 0) outperformed MC (3 ± 1), with a significant difference (**p = 0.008**), reinforcing ChatGPT’s superiority in delivering informative responses (Table 2).

Evaluation of AI Models in Answering “How Often Should Short Stature Be Monitored?”

In the “All Necessary Information Provided” category, MC scored the lowest (2 ± 1), while ChatGPT and Bard performed better (3 ± 1). A significant difference was observed between MC and ChatGPT (**p = 0.018**), indicating MC’s weaker performance in delivering comprehensive information. In “Missing Information Provided” MC (4 ± 1) had the highest score, showing a greater tendency to omit details, with a significant difference from ChatGPT (**p = 0.018**). In “Medically Accurate Information Provided” ChatGPT (4 ± 1) outperformed MC (3 ± 0), with a statistically significant difference (**p = 0.013**), highlighting ChatGPT’s superior accuracy. For “Recommendations Provided” Bard (4 ± 1) and ChatGPT (4 ± 0) led, while Pi (3 ± 1) performed significantly worse (**p < 0.001**). In “Response Was Sufficient for the Patient” ChatGPT and Bard (4 ± 1) excelled, while MC (2 ± 1) and Pi (3 ± 1) scored lower. A significant difference was found between ChatGPT and MC (**p = 0.003**), emphasizing ChatGPT’s stronger ability to meet users' informational needs. Lastly, in “Recommendation to Consult a Physician” Bard (4 ± 0) and ChatGPT (4 ± 1) performed best, while Pi (2 ± 1) had the lowest score, with significant differences between Pi and the other models (**p < 0.001**) (Table 2).

Evaluation of AI Models in Answering “What Are the Side Effects That Can Be Seen After Growth Hormone Treatment?”

In the “Excessive Information Provided” category, MC (4 ± 1) had the highest score, significantly differing from ChatGPT (2 ± 1) and Pi (2 ± 1) (**p = 0.015**), indicating MC’s greater tendency to provide excessive details. In “Recommendations Provided” Bard (4 ± 0) and ChatGPT (4 ± 1) scored the highest, while Pi (2 ± 1) performed the worst. A significant difference was observed between Pi and the other models (**p = 0.001**), suggesting Pi’s limitations in providing recommendations. For “Patient Guidance Provided” Bard (4 ± 0) and ChatGPT (4 ± 1) again excelled, while Pi (2 ± 1) lagged significantly (**p < 0.001**), demonstrating Bard and ChatGPT’s superior ability to offer guidance. In “Recommendation to Consult a Physician” Bard (4 ± 1) and ChatGPT (4 ± 1) performed best, while Pi (2 ± 0) had the lowest score. Post-hoc analysis confirmed Pi’s significantly weaker performance compared to the other models (**p < 0.001**), reinforcing Bard and ChatGPT’s reliability in clinical guidance. Lastly, in “Response Aims to Inform the Reader” Bard (4 ± 0) and ChatGPT (4 ± 1) scored the highest, while Pi (3 ± 1) performed worse, with a significant difference between Pi and ChatGPT (**p = 0.030**), highlighting ChatGPT’s strength in delivering informative responses (Table 2).

Table 3 presents the expert evaluation of responses provided by different AI models (ChatGPT, Bard, MC, and Pi) to questions related to pediatric endocrinology.

A statistically significant difference was found among AI models for the “Was a definition provided?” question (**p = 0.028**). The observed differences were between MC-ChatGPT and Pi-ChatGPT. For the “Was all necessary information provided?” question, Bard (3.5 ± 0.4) had the highest score, while MC (2.7 ± 0.5) had the lowest average. A statistically significant difference was detected among the AI models (**p = 0.002**),

UNCORRECTED PROOF

UNCORRECTED PROOF

with differences specifically identified between MC-ChatGPT and MC-Bard. For the "Was missing information provided?", "Was excessive information provided?", "Was irrelevant information provided?", and "Was the information medically accurate?" questions, the respective p values were 0.074, 0.178, 0.486, and 0.12, indicating no statistically significant differences among the AI models ($p > 0.05$). In the "Were recommendations provided?" question, Bard (4 ± 0.3) had the highest score, while Pi (2.4 ± 0.6) had the lowest. A statistically significant difference was observed among the AI models (**p = 0.001**), with significant differences identified between MC-Bard and Pi-Bard. Similarly, in the "Was patient guidance provided?" criterion, Bard (4.1 ± 0.3) had the highest average score, while Pi (2.3 ± 0.6) had the lowest. A statistically significant difference was found among the AI models (**p < 0.001**), with the difference primarily between Pi and Bard. For the "Was a recommendation to consult a physician provided?" question, Bard (4.2 ± 0.5) received the highest score, and a significant difference was identified between Pi and Bard (**p < 0.001**). For the "Was the response sufficient for the patient?" question, Bard (3.5 ± 0.3) and ChatGPT (3.4 ± 0.5) had similar values, receiving the highest scores. A statistically significant difference was found among the AI models (**p = 0.004**), with differences observed between MC-ChatGPT and MC-Bard. A statistically significant difference was also detected for the "Does the response aim to inform the reader?" question (**p = 0.045**), with the difference identified between Pi and ChatGPT. Finally, for the "Does the response aim to reassure the reader?" question, ChatGPT (2.8 ± 0.2) had the highest value, and a statistically significant difference was found among the AI models (**p = 0.007**). The observed differences were between MC-ChatGPT and Pi-ChatGPT. The highest reliability, with an ICC value of 0.774 (0.682–0.844), was observed in the question "Was an appropriate definition provided?", while the lowest reliability, with an ICC value of -0.047 (-0.306–0.197), was observed in the question "Was any recommendation given?" (Table 3).

Discussion

In this study, responses provided by four different AI models (ChatGPT, Bard, MC, and Pi) to questions related to pediatric endocrinology were evaluated by experts based on specific criteria. The findings indicate significant differences among AI models in terms of medical information accuracy, guidance capacity, and user informativeness. The Bard model distinguished itself in categories requiring guidance and direction (e.g., "Were recommendations provided?" and "Was a recommendation to consult a physician provided?"), receiving the highest scores. Additionally, Bard was the model that provided the least missing information. This success suggests that Bard possesses a strong ability to deliver supportive and guiding responses. The literature highlights that AI models focusing on guidance enhance user confidence and support medical decision-making processes (7). MC excelled in providing accurate and medically reliable information. In categories such as "Was medically accurate information provided?" and "Was all necessary information provided?", it performed similarly to or even outperformed Bard and ChatGPT. This indicates that MC is a reliable model for areas requiring medical accuracy. However, its lower guidance capacity suggests that it may not be sufficient for clinical applications on its own. Existing literature also supports the notion that AI models with strong medical accuracy capabilities are effectively utilized in clinical decision support systems (8). ChatGPT demonstrated consistent performance in providing information and educating users. It received high scores in the "Does the response aim to inform the reader?" category, highlighting its reliability as an informational model. However, it lagged behind Bard and MC in categories requiring guidance. This finding suggests that while ChatGPT is effective in knowledge dissemination, it requires further development in user guidance. AI models with strong user education capabilities are known to play a crucial role in patient education and information dissemination (9). The Pi model exhibited acceptable performance in basic informational categories but received the lowest scores in guidance and recommendation-based categories. This suggests that Pi is inadequate for guidance-focused clinical decision-making processes. AI models with limited guidance capacities are generally considered more suitable for handling basic queries rather than facilitating detailed information sharing (10). Overall, Bard emerges as the most effective model in terms of guidance and recommendations, making it a more suitable AI for specialized fields such as pediatric endocrinology, where expert guidance is essential. MC is a medically accurate model, but it requires improvement in its guidance capabilities. ChatGPT demonstrates strong informational capabilities, and if it enhances its guidance capacity, it could have broader applications. Meanwhile, Pi shows significant limitations in guidance and recommendations, making it insufficient for clinical applications requiring decision support.

The findings of this study highlight the strengths and weaknesses of different AI models and shed light on their potential applications in medical decision-making processes. For instance, Bard, with its strong guidance capacity, could be beneficial in patient management, while MC may be more effective in areas that require medical accuracy. ChatGPT stands out as a suitable model for patient education and general information sharing.

Study Limitations

The answers produced may change due to the updating of the AI models used in our study. The answers of the experts making the evaluations may contain subjectivity. The lack of real patient data in our study can be considered as a limitation. More studies are needed for the integration of AI in clinical applications. The fact that AI platforms are subject to change over time and are constantly evolving may lead to differences in the results of the study if the same analysis is performed at a later date.

Conclusion

This study demonstrated that different AI models exhibit varying performances in the field of pediatric endocrinology. The Bard model excelled in guidance and recommendation categories, while MC proved to be strong in medical accuracy. ChatGPT emerged as a reliable option for information dissemination and user education, whereas Pi showed limited applicability in this domain due to its insufficient guidance capacity. Future research should focus on improving AI models to achieve a more balanced performance in both guidance and medical accuracy. Additionally, optimizing these models to align with user needs is recommended to enhance patient trust and integrate AI effectively into clinical decision-support processes. The literature emphasizes that while AI holds great potential in supporting patient care processes, this potential can only be fully realized through a careful balance in model design (11,12). These findings represent a significant step toward enhancing AI's role in clinical applications and underscore the need for developing customized AI solutions tailored to the needs of specialized fields such as pediatric endocrinology.

Ethics: This study does not require ethical committee approval.

Informed Consent: Mixed-methods.

Footnotes

Authorship Contributions

Concept: K.K., O.F.O., A.C. Design: K.K., O.F.O., M.U. Data Collection or Processing: K.K., A.S.D., S.K.K., E.D. Analysis or Interpretation: K.K., O.F.O., A.C. Literature Search: K.K., O.F.O., U.O. Writing: K.K., A.C., H.D.

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

References

1. Demirel, O., & Sonuç, E. (2021). Yapay Zeka Teknikleri Kullanılarak Kemik Yaşı Tespiti. Türkiye Sağlık Enstitüleri Başkanlığı Dergisi, 4(3), 17-30. https://doi.org/10.54537/tusebdergisi.1023666

2. Ferrante, G., Licari, A., Fasola, S., Marseglia, G. L., & La Grutta, S. (2021). Artificial intelligence in the diagnosis of pediatric allergic diseases. *Pediatric allergy and immunology : official publication of the European Society of Pediatric Allergy and Immunology*, *32*(3), 405–413. https://doi.org/10.1111/pai.13419

3. Zhang, L., Chen, J., Hou, L., Xu, Y., Liu, Z., Huang, S., Ou, H., Meng, Z., & Liang, L. (2022). Clinical application of artificial intelligence in longitudinal image analysis of bone age among GHD patients. *Frontiers in Pediatrics*, 10. https://doi.org/10.3389/fped.2022.986500

4. Winkelman, J., Nguyen, D., vanSonnenberg, E., Kirk, A., & Lieberman, S. (2023). Artificial Intelligence (AI) in pediatric endocrinology. *Journal of Pediatric Endocrinology and Metabolism*, 36, 903-908. https://doi.org/10.1515/jpem-2023-0287

5. Otjen, J. P., Moore, M. M., Romberg, E. K., Perez, F., & Iyer, R. (2021). The current and future roles of artificial intelligence in pediatric radiology. *Pediatric Radiology*, 52, 2065-2073. https://doi.org/10.1007/s00247-021-05086-9

6. Waikel, R., Othman, A. A., Patel, T., Ledgister Hanchard, S. L., Hu, P., Tekendo-Ngongang, C., Van Duong, D., & Solomon, B. D. (2023). Generative Artificial Intelligence Methods for Pediatric Genetics Education. *medRxiv*. https://doi.org/10.1101/2023.08.01.23293506

7. Karalis, V. D. (2024). The Integration Of Artificial Intelligence Into Clinical Practice. *Applied Biosciences*, *3*(1), 14-44. Https://Doi.Org/10.3390/Applbiosci3010002

8. Nyiramana Mukamurera P. (2024). The Role Of Artificial Intelligence In Clinical Decision Support Systems. Research Invention Journal Of Public Health And Pharmacy 3(2):14-17. https://Doi.Org/10.59298/RIJPP/2024/321417

9. Guo F, Zhou A, Chang W, Sun X and Zou B (2022) Is physician online information sharing always beneficial to patient education? An attention perspective. Front. Public Health 10:987766. doi: 10.3389/fpubh.2022.987766

10. Bekbolatova, M., Mayer, J., Ong, C. W., & Toma, M. (2024). Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. *Healthcare (Basel, Switzerland)*, *12*(2), 125. https://doi.org/10.3390/healthcare12020125

11. Kranenburg, L., Reerds, S., Cools, M., Alderson, J., Muscarella, M., Magrite, E., Kuiper, M., Abdelgaffar, S., Balsamo, A., Brauner, R., Chanoine, J., Deeb, A., Fechner, P., German, A., Holterhus, P., Juul, A., Mendonca, B., Neville, K., Nordenstrom, A., Oostdijk, W., Rey, R., Rutter, M., Shah, N., Luo, X., Grijpink, K., & Drop, S. (2017). Global Application of the Assessment of Communication Skills of Paediatric Endocrinology Fellows in the Management of Differences in Sex Development Using the ESPE E-Learning.Org Portal. Hormone Research in Paediatrics, 88, 127 - 139. https://doi.org/10.1159/000475992

12. Kalaitzoglou, E., Majaliwa, E., Zacharin, M., De Beaufort, C., Chanoine, J., Van Wijngaard-DeVugt, C., Sperla, E., Boot, A., & Drop, S. (2020). Multilingual Global E-Learning Pediatric Endocrinology and Diabetes Curriculum for Front Line Health Care Providers in Resource-Limited Countries: Development Study. JMIR Formative Research, 4. https://doi.org/10.2196/18555

Table 1. Expert Evaluation of AI-Generated Responses to Pediatric Endocrinology Questions I

		Group						
		ChatGPT	Bard	MC	Pi			
		Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Kruskal-Wallis H	p	post-hoc
What is short stature?	Was a proper definition provided?	4 ± 1; 4±2	5 ± 1; 5±1	4 ± 1; 4±1	4 ± 1; 4±0	5.498	0.139	
	Was all necessary information included?	3 ± 1; 2.5±2	4 ± 1; 4±1	3 ± 1; 3±1	3 ± 1; 3±1	4.707	0.195	
	Was any essential information missing?	4 ± 1; 4±0	3 ± 1; 3±2	3 ± 1; 3±1	3 ± 1; 4±2	9.150	0.027	Bard-ChatGPT
	Was excessive information provided?	2 ± 1; 1.5±2	2 ± 1; 2±3	2 ± 1; 2±1	2 ± 1; 2±1	1.558	0.669	
	Was any irrelevant information included?	2 ± 1; 2±1	3 ± 2; 3±3	2 ± 1; 2±2	2 ± 1; 2±3	0.650	0.885	
	Was the medical information accurate?	3 ± 1; 2.5±3	4 ± 1; 3.5±1	3 ± 1; 3±2	3 ± 1; 3±1	3.289	0.349	
	Were recommendations given?	2 ± 1; 2±2	4 ± 0; 4±0	4 ± 1; 4±1	2 ± 1; 2±1	29.005	**<0.001**	Pi-MC, Pi-Bard, ChatGPT-MC, ChatGPT-Bard
	Was patient guidance provided?	2 ± 1; 2±1	4 ± 1; 4±1	4 ± 0; 4±0	2 ± 1; 1.5±1	26.564	**<0.001**	Pi-MC, Pi-Bard, ChatGPT-MC, ChatGPT-Bard
	Was a recommendation to consult a physician included?	2 ± 1; 2±1	4 ± 1; 4±1	4 ± 0; 4±0	1 ± 0; 1±1	30.593	**<0.001**	Pi-MC, Pi-Bard, ChatGPT-MC, ChatGPT-Bard
	Was the response sufficient for the patient?	2 ± 1; 2±1	3 ± 1; 4±2	3 ± 1; 3±2	3 ± 1; 3±2	6.923	0.074	
	Did the response aim to inform the reader?	3 ± 1; 3±2	4 ± 1; 4±1	4 ± 0; 4±0	3 ± 1; 3±1	12.160	**0.007**	ChatGPT-MC
	Did the response aim to reassure the reader?	3 ± 1; 3±1	4 ± 1; 4±2	3 ± 1; 2±2	2 ± 1; 2±0	18.140	**<0.001**	Pi-Bard, MC-Bard
What are the causes of	Was a proper definition provided?	4 ± 1; 4±3	3 ± 1; 4±2	4 ± 1; 3.5±1	3 ± 1; 2±2	5.137	0.162	
	Was all necessary information included?	4 ± 1; 4±1	3 ± 1; 4±2	2 ± 0; 2±0	2 ± 1; 2±1	15.588	**0.001**	MC-ChatGPT, Pi-ChatGPT

UNCORRECTED PROOF

short stature ?	Was any essential information missing?	3 ± 1; 2.5±1	2 ± 1; 2±3	4 ± 1; 4±1	4 ± 1; 4±0	11.076	**0.011**	Bard-Pi
	Was excessive information provided?	2 ± 1; 2±0	2 ± 1; 2±2	3 ± 1; 2±2	2 ± 0; 2±1	5.013	0.171	
	Was any irrelevant information included?	2 ± 0; 2±1	2 ± 1; 2±1	3 ± 1; 3±1	2 ± 1; 2±2	11.176	**0.011**	Bard-MC, ChatGPT-MC
	Was the medical information accurate?	4 ± 1; 4±2	4 ± 1; 4±1	3 ± 1; 3±1	3 ± 1; 3.5±2	11.114	**0.011**	MC-ChatGPT
	Were recommendations given?	3 ± 1; 3.5±2	4 ± 1; 4±1	3 ± 1; 3±1	2 ± 1; 2±0	23.636	**<0.001**	Pi-MC, Pi-Bard
	Was patient guidance provided?	3 ± 1; 3±2	5 ± 0; 5±1	4 ± 1; 4±1	2 ± 1; 2±1	24.831	**<0.001**	Pi-MC, Pi-Bard, ChatGPT-Bard
	Was a recommendation to consult a physician included?	3 ± 1; 3±2	5 ± 1; 5±1	4 ± 1; 4±1	2 ± 1; 2±1	23.756	**<0.001**	Pi-MC, Pi-Bard
	Was the response sufficient for the patient?	4 ± 1; 4±2	4 ± 1; 4±1	3 ± 1; 3±1	2 ± 1; 2±2	20.325	**<0.001**	Pi-ChatGPT, Pi-Bard
	Did the response aim to inform the reader?	4 ± 1; 4±1	4 ± 1; 4±0	4 ± 1; 4±1	3 ± 1; 3±2	8.278	**0.041**	Pi-Bard
	Did the response aim to reassure the reader?	3 ± 1; 3±1	4 ± 2; 4.5±3	3 ± 1; 2±1	3 ± 1; 3±2	4.135	0.247	
How is growth rate evaluated in short stature ?	Was a proper definition provided?	4 ± 1; 4±1	3 ± 2; 3±4	3 ± 1; 2±2	3 ± 1; 2±1	8.433	**0.038**	Pi-ChatGPT
	Was all necessary information included?	3 ± 1; 4±2	4 ± 1; 4±1	3 ± 1; 3±1	4 ± 1; 4±1	5.875	0.118	
	Was any essential information missing?	3 ± 1; 2±3	2 ± 1; 2±1	4 ± 1; 4±1	2 ± 1; 2±1	8.412	**0.038**	Bard-MC
	Was excessive information provided?	2 ± 1; 2±0	3 ± 1; 2±2	2 ± 1; 2±1	2 ± 0; 2±1	4.264	0.234	
	Was any irrelevant information included?	2 ± 1; 2±1	2 ± 1; 2±3	2 ± 1; 2±0	3 ± 1; 2±3	1.124	0.771	
	Was the medical information accurate?	4 ± 1; 4±1	4 ± 1; 3±1	4 ± 0; 4±1	4 ± 1; 4±0	3.487	0.322	
	Were recommendations given?	3 ± 1; 2.5±2	4 ± 0; 4±0	2 ± 1; 1.5±1	3 ± 1; 2.5±2	18.720	**<0.001**	ChatGPT-Bard, MC-Bard
	Was patient guidance provided?	3 ± 1; 3.5±1	4 ± 1; 4±2	2 ± 1; 1.5±1	3 ± 1; 2±2	12.110	**0.007**	MC-Bard
	Was a recommendation to consult a physician included?	4 ± 1; 4±1	4 ± 1; 5±2	2 ± 1; 2±1	3 ± 1; 3±2	13.789	**0.003**	MC-Bard
	Was the response sufficient for the patient?	4 ± 1; 4±3	4 ± 1; 4±2	2 ± 1; 2±0	4 ± 1; 4±2	8.428	**0.038**	MC-ChatGPT, MC-Bard, MC-Pi
	Did the response aim to inform the reader?	4 ± 0; 4±1	4 ± 1; 4±1	3 ± 1; 3±1	4 ± 1; 4±1	9.917	**0.019**	MC-Bard
	Did the response aim to reassure the reader?	3 ± 1; 3.5±2	3 ± 2; 4±3	2 ± 1; 2±2	3 ± 1; 2.5±2	4.339	0.227	
How is bone age determined in short stature ?	Was a proper definition provided?	4 ± 1; 4±1	3 ± 1; 4±2	3 ± 1; 4±2	3 ± 1; 4±2	2.801	0.423	
	Was all necessary information included?	4 ± 1; 3.5±2	4 ± 1; 4±1	3 ± 1; 3±1	3 ± 1; 3±1	6.513	0.089	
	Was any essential information missing?	3 ± 1; 3±3	3 ± 1; 3±2	4 ± 1; 4±1	3 ± 1; 4±1	6.894	0.075	
	Was excessive information provided?	2 ± 1; 2±0	3 ± 1; 2±2	3 ± 1; 3.5±2	2 ± 0; 2±1	5.323	0.150	
	Was any irrelevant information included?	2 ± 1; 2.5±1	3 ± 1; 2±1	3 ± 1; 3±2	2 ± 1; 2±1	4.813	0.186	
	Was the medical information accurate?	4 ± 1; 4±0	3 ± 1; 3±1	3 ± 1; 3±2	3 ± 1; 3±1	6.781	0.079	
	Were recommendations given?	3 ± 1; 3.5±1	4 ± 1; 4±1	2 ± 1; 2±1	2 ± 1; 1.5±2	18.220	**<0.001**	Pi-Bard, MC-Bard
	Was patient guidance provided?	3 ± 1; 3±1	4 ± 0; 4±1	2 ± 1; 2.5±2	2 ± 1; 2±1	19.372	**<0.001**	Pi-Bard, ChatGPT-Bard, MC-Bard
	Was a recommendation to consult a physician included?	4 ± 1; 3.5±2	5 ± 1; 4.5±1	2 ± 0; 2±1	2 ± 1; 2±0	25.323	**<0.001**	Pi-Bard, MC-Bard, MC-ChatGPT
	Was the response sufficient for the patient?	3 ± 1; 3.5±1	4 ± 1; 4±1	3 ± 1; 3±1	3 ± 1; 3±1	8.178	**0.042**	MC-Bard
	Did the response aim to inform the reader?	4 ± 0; 4±0	4 ± 1; 4±1	3 ± 1; 3±1	4 ± 1; 4±0	7.096	0.069	
	Did the response aim to reassure the reader?	3 ± 1; 2.5±2	2 ± 1; 2±1	2 ± 1; 2±1	2 ± 1; 2±1	5.522	0.137	

UNCORRECTED PROOF

What should be considered in differential diagnosis in short stature?	Was a proper definition provided?	4 ± 1; 4±1	3 ± 1; 2.5±2	4 ± 1; 4±1	3 ± 1; 3±2	8.321	**0.040**	Bard-ChatGPT
	Was all necessary information included?	4 ± 1; 4±2	3 ± 1; 4±1	3 ± 1; 3.5±2	3 ± 1; 3±1	1.606	0.658	
	Was any essential information missing?	3 ± 2; 3±3	3 ± 1; 3±2	2 ± 1; 2±2	3 ± 1; 4±2	4.436	0.218	
	Was excessive information provided?	2 ± 1; 2±2	2 ± 1; 2±1	2 ± 1; 2±0	2 ± 0; 2±1	2.172	0.537	
	Was any irrelevant information included?	2 ± 1; 1.5±1	2 ± 0; 2±0	2 ± 0; 2±1	2 ± 1; 2±2	1.461	0.691	
	Was the medical information accurate?	4 ± 1; 4±2	4 ± 0; 4±0	4 ± 1; 4±1	3 ± 1; 4±2	2.303	0.512	
	Were recommendations given?	4 ± 1; 3±1	4 ± 1; 4±2	4 ± 0; 4±0	3 ± 1; 3±2	6.707	0.082	
	Was patient guidance provided?	3 ± 1; 2±2	4 ± 0; 4±1	4 ± 0; 4±0	3 ± 1; 3±2	12.659	**0.005**	Pi-Bard, ChatGPT-Bard
	Was a recommendation to consult a physician included?	3 ± 1; 3±2	5 ± 1; 5±1	5 ± 1; 4.5±1	3 ± 1; 2.5±3	15.615	**0.001**	Pi-MC, Pi-Bard
	Was the response sufficient for the patient?	4 ± 1; 4±2	4 ± 1; 4±1	4 ± 1; 4±1	3 ± 1; 2.5±1	7.403	0.060	
	Did the response aim to inform the reader?	4 ± 1; 4±1	4 ± 1; 4±0	4 ± 0; 4±0	3 ± 1; 3±1	13.163	**0.004**	Pi-ChatGPT
	Did the response aim to reassure the reader?	3 ± 1; 2±2	2 ± 1; 2±3	3 ± 1; 2±1	2 ± 1; 2±0	0.598	0.897	

Table 2. Expert Evaluation of AI-Generated Responses to Pediatric Endocrinology Questions II

		Group					
		ChatGPT	Bard	MC	Pi		
		Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Mean ± SD; Med ± IQR	Kruskal-Wallis H	p
What should be considered in laboratory parameters in short stature?	Was a proper definition provided?	4 ± 2; 4±3	3 ± 1; 3±2	3 ± 1; 2±2	3 ± 1; 3±2	1.894	0.595
	Was all necessary information included?	4 ± 1; 3.5±3	3 ± 1; 3±1	3 ± 1; 3±2	3 ± 1; 4±1	2.669	0.446
	Was any essential information missing?	3 ± 2; 3±3	3 ± 1; 3±1	3 ± 1; 3.5±3	3 ± 1; 3±2	1.189	0.756
	Was excessive information provided?	2 ± 1; 2±3	2 ± 1; 2±1	3 ± 1; 2±2	3 ± 1; 2±3	1.543	0.672
	Was any irrelevant information included?	3 ± 1; 3.5±2	2 ± 1; 2±1	3 ± 1; 4±2	3 ± 1; 2±3	3.593	0.309
	Was the medical information accurate?	3 ± 1; 4±2	3 ± 1; 3.5±2	3 ± 0; 3±0	3 ± 1; 3±2	0.598	0.897
	Were recommendations given?	3 ± 1; 3±1	3 ± 1; 3±1	2 ± 1; 2±2	3 ± 1; 3±1	5.295	0.151
	Was patient guidance provided?	3 ± 1; 3±2	4 ± 1; 4±1	2 ± 1; 2±1	3 ± 1; 3±2	8.962	**0.030**
	Was a recommendation to consult a physician included?	3 ± 1; 2±2	3 ± 1; 4±2	2 ± 1; 1±1	3 ± 1; 3±2	10.588	**0.014**
	Was the response sufficient for the patient?	4 ± 1; 4±2	3 ± 1; 4±2	2 ± 1; 2±1	3 ± 1; 3.5±1	10.111	**0.018**
	Did the response aim to inform the reader?	4 ± 1; 4±1	4 ± 1; 4±1	3 ± 1; 3±1	4 ± 0; 4±0	8.726	**0.033**
	Did the response aim to reassure the reader?	3 ± 1; 2±2	2 ± 1; 2±1	3 ± 1; 2.5±2	3 ± 1; 2±1	0.944	0.815
Which drugs are used in the treatment of short stature?	Was a proper definition provided?	4 ± 2; 4.5±3	3 ± 2; 4±3	2 ± 1; 2±1	3 ± 1; 3±2	6.971	0.073
	Was all necessary information included?	4 ± 1; 4±2	4 ± 1; 4±1	2 ± 1; 2±0	3 ± 1; 2±2	19.178	**<0.001**
	Was any essential information missing?	2 ± 1; 2±2	3 ± 1; 3±2	4 ± 1; 4±2	2 ± 1; 2±1	15.036	**0.002**
	Was excessive information provided?	2 ± 1; 1.5±3	3 ± 2; 2±3	2 ± 0; 2±1	2 ± 1; 2±2	2.103	0.551
	Was any irrelevant information included?	2 ± 1; 2±1	2 ± 1; 2±1	2 ± 1; 2±2	2 ± 1; 2±2	0.600	0.896
	Was the medical information accurate?	4 ± 1; 4±1	3 ± 1; 4±2	3 ± 1; 2.5±1	3 ± 0; 3±1	8.565	**0.036**
	Were recommendations given?	3 ± 1; 3.5±1	4 ± 1; 4±1	3 ± 1; 3±1	3 ± 1; 2.5±2	9.182	**0.027**
	Was patient guidance provided?	3 ± 1; 3.5±1	4 ± 1; 4±0	3 ± 1; 3±1	3 ± 1; 2.5±2	11.977	**0.007**
	Was a recommendation to consult a physician included?	3 ± 1; 4±2	3 ± 2; 4±3	3 ± 1; 4±2	3 ± 1; 3±0	3.347	0.341
	Was the response sufficient for the patient?	4 ± 1; 4±1	3 ± 1; 3±2	2 ± 1; 2±1	2 ± 1; 2±2	14.788	**0.002**
	Did the response aim to inform the reader?	4 ± 0; 4±0	3 ± 1; 4±1	3 ± 1; 3±2	4 ± 1; 3.5±1	11.880	**0.008**
	Did the response aim to reassure the reader?	3 ± 1; 3±2	2 ± 1; 2±2	2 ± 1; 2±1	2 ± 1; 2±1	10.577	**0.014**
How often should short stature be monitored?	Was a proper definition provided?	3 ± 1; 4±2	3 ± 1; 2±2	2 ± 1; 2±0	3 ± 1; 3±2	6.086	0.108
	Was all necessary information included?	3 ± 1; 3.5±1	3 ± 1; 4±2	2 ± 1; 2±0	3 ± 1; 3±1	10.044	**0.018**
	Was any essential information missing?	3 ± 1; 2.5±3	3 ± 1; 3±0	4 ± 1; 4±0	3 ± 1; 4±2	10.087	**0.018**
	Was excessive information provided?	2 ± 1; 2±1	2 ± 1; 2±2	2 ± 0; 2±1	2 ± 1; 2±1	1.419	0.701
	Was any irrelevant information included?	2 ± 1; 2±2	2 ± 0; 2±1	3 ± 2; 2.5±4	3 ± 1; 2±3	3.018	0.389
	Was the medical information accurate?	4 ± 1; 4±0	3 ± 1; 3±1	3 ± 0; 3±0	4 ± 1; 3.5±1	10.702	**0.013**
	Were recommendations given?	4 ± 1; 4±0	4 ± 0; 4±0	3 ± 1; 3±1	3 ± 1; 2±1	18.475	**<0.001**
	Was patient guidance provided?	4 ± 1; 4±0	4 ± 0; 4±0	4 ± 1; 4±1	2 ± 1; 2±1	14.169	**0.003**
	Was a recommendation to consult a physician included?	4 ± 0; 4±0	4 ± 1; 4.5±1	4 ± 0; 4±1	2 ± 1; 2±0	20.644	**<0.001**

UNCORRECTED PROOF

	Was the response sufficient for the patient?	4 ± 1; 4±1	4 ± 1; 4±1	2 ± 1; 2±0	3 ± 1; 2.5±1	13.867	**0.003**
	Did the response aim to inform the reader?	4 ± 1; 4±1	4 ± 0; 4±0	4 ± 1; 3.5±1	3 ± 1; 3±2	6.194	0.103
	Did the response aim to reassure the reader?	3 ± 1; 2±2	3 ± 1; 3±3	2 ± 1; 2±0	2 ± 1; 2±1	4.128	0.248
What are the side effects that can be seen after growth hormone treatment?	Was a proper definition provided?	3 ± 2; 4±3	3 ± 1; 2.5±3	2 ± 1; 2±2	3 ± 1; 2±2	2.667	0.446
	Was all necessary information included?	3 ± 1; 3±2	3 ± 1; 2.5±2	3 ± 1; 3±2	2 ± 1; 2±1	2.281	0.516
	Was any essential information missing?	3 ± 1; 2.5±3	3 ± 1; 3±2	3 ± 1; 2±2	3 ± 1; 4±2	1.942	0.585
	Was excessive information provided?	2 ± 1; 2±1	3 ± 1; 4±2	4 ± 1; 4±0	2 ± 1; 1.5±3	10.448	**0.015**
	Was any irrelevant information included?	2 ± 1; 2±2	3 ± 1; 2±2	3 ± 1; 3±1	2 ± 1; 2±2	5.300	0.151
	Was the medical information accurate?	4 ± 1; 4±1	3 ± 1; 3±2	3 ± 1; 3±2	3 ± 1; 2±2	5.665	0.129
	Were recommendations given?	4 ± 0; 4±0	4 ± 1; 4±2	3 ± 1; 3.5±2	2 ± 1; 2±1	17.006	**0.001**
	Was patient guidance provided?	4 ± 0; 4±0	4 ± 1; 4±2	3 ± 1; 3.5±2	2 ± 1; 2±1	17.790	**<0.001**
	Was a recommendation to consult a physician included?	4 ± 1; 4±1	4 ± 1; 4±1	3 ± 1; 3.5±2	2 ± 0; 2±1	22.334	**<0.001**
	Was the response sufficient for the patient?	3 ± 1; 4±2	3 ± 1; 3±2	3 ± 1; 3±2	2 ± 1; 2±2	7.628	0.054
	Did the response aim to inform the reader?	4 ± 0; 4±0	4 ± 1; 4±1	4 ± 1; 4±0	3 ± 1; 3±1	8.965	**0.030**
	Did the response aim to reassure the reader?	3 ± 1; 2.5±2	2 ± 1; 2±1	2 ± 1; 1.5±1	2 ± 1; 2±1	6.925	0.074

Table 3. Comparison of Expert Evaluation Averages for AI-Generated Responses to Questions

Questions	**ChatGPT**	**Bard**	**MC**	**Pi**				
	Mean ± SD; Median (Min-Max)	Mean ± SD; Median (Min-Max)	Mean ± SD; Median (Min-Max)	Mean ± SD; Median (Min-Max)	**Kruskal-Wallis H**	**p**	**Post-hoc**	**ICC; 95% CI (L -U)**
Was a proper definition provided?	3.7 ± 0.3; 3.6 (3.3-4.1)	3.2 ± 0.6; 3.1 (2.7-4.5)	3 ± 0.8; 2.7 (2.1-4)	3.1 ± 0.5; 3.1 (2.5-4.1)	9.066	**0.028**	MC-ChatGPT, Pi-ChatGPT	0.774 (0.682 - 0.844)
Was all necessary information included?	3.4 ± 0.3; 3.5 (2.8-3.8)	3.5 ± 0.4; 3.4 (2.9-3.9)	2.7 ± 0.5; 2.8 (1.9-3.3)	2.9 ± 0.5; 3 (2.2-3.5)	14.596	**0.002**	MC-ChatGPT, MC-Bard,	0.523 (0.343 - 0.664)
Was any essential information missing?	2.7 ± 0.5; 2.6 (2.2-3.9)	2.8 ± 0.3; 2.9 (2.3-3.1)	3.3 ± 0.7; 3.6 (2.2-4)	3 ± 0.5; 3.2 (2-3.8)	6.922	0.074		0.611 (0.463 - 0.726)
Was excessive information provided?	2.1 ± 0.2; 2.1 (1.8-2.3)	2.4 ± 0.4; 2.4 (1.9-3)	2.4 ± 0.8; 2.2 (1.7-4)	2 ± 0.3; 1.9 (1.7-2.5)	4.912	0.178		0.525 (0.345 - 0.666)
Was any irrelevant information included?	2.3 ± 0.4; 2.4 (1.7-3)	2.2 ± 0.4; 2.2 (1.6-2.8)	2.6 ± 0.6; 2.7 (1.7-3.3)	2.3 ± 0.2; 2.4 (1.8-2.5)	2.443	0.486		0.536 (0.360 - 0.674)
Was the medical information accurate?	3.6 ± 0.4; 3.8 (2.8-4)	3.4 ± 0.4; 3.4 (3-4)	3.1 ± 0.4; 3.1 (2.5-3.8)	3.7 ± 1.4; 3.3 (2.6-7.4)	5.842	0.12		0.444 (0.237 - 0.607)
Were recommendations given?	3.3 0.6; 3.3 (2.1-4.1)	4 0.3; 4 (3.2-4.4)	2.9 0.8; 2.8 (1.8-4.1)	2.4 0.6; 2.5 (1.6-3.2)	17.659	**0.001**	MC-Bard, Pi-Bard,	-0.047 (-0.306-0.197)

UNCORRECTED PROOF

Was patient guidance provided?	3.2 0.6; 3.1 (2.3-4.1)	4.1 0.3; 4 (3.6-4.7)	3.1 0.9; 3.4 (1.8-4.2)	2.3 0.6; 2.3 (1.5-3.1)	18.498	**<0.001**	Pi- Bard,	-0.030 (-0.272-0.202)
Was a recommendation to consult a physician included?	3.2 0.6; 3.4 (1.9-4.1)	4.2 0.5; 4.4 (3.4-4.6)	3.1 1.1; 3.2 (1.5-4.5)	2.2 0.6; 2.1 (1.3-3.1)	17.992	**<0.001**	Pi-Bard,	0.049 (-0.181 - 0.268)
Was the response sufficient for the patient?	3.4 0.5; 3.5 (2.1-3.9)	3.5 0.3; 3.6 (3-4)	2.6 0.6; 2.5 (1.8-3.7)	2.8 0.6; 2.7 (2-3.6)	13.305	**0.004**	MC-ChatGPT, MC-Bard,	0.341 (0.112 - 0.527)
Did the response aim to inform the reader?	3.9 0.4; 4.1 (2.9-4.4)	3.9 0.3; 4 (3.4-4.4)	3.6 0.4; 3.5 (2.8-4.2)	3.5 0.4; 3.4 (3-4.1)	8.048	**0.045**	Pi-ChatGPT,	0.364 (0.135 - 0.548)
Did the response aim to reassure the reader?	2.8 0.2; 2.8 (2.5-3.1)	2.8 0.7; 2.4 (2.3-4.1)	2.2 0.4; 2.2 (1.6-2.7)	2.3 0.4; 2 (1.8-2.9)	12.059	**0.007**	MC-ChatGPT, Pi-ChatGPT	0.523 (0.345 - 0.663)

UNCORRECTED PROOF