RESEARCH ARTICLE

F.U. Med.J.Health.Sci.
2026; 40 (1): 98 - 104
http://www.fusabil.org

Comparative Evaluation of Three Artificial Intelligence Chatbots in Providing Information on Pediatric Celiac Disease

Ecem İpek ALTINOK [1, a]
Özlem SÜMER COŞAR [2, b]
Volkan ALTINOK [3, c]

[1] Ordu University, Faculty of Medicine, Department of Child Health and Diseases, Ordu, TÜRKİYE

[2] Gazi University, Faculty of Medicine, Department of Pediatric Gastroenterology, Ankara, TÜRKİYE

[3] Ordu University, Faculty of Medicine, Department of Pediatric Surgery, Ordu, TÜRKİYE

[a] ORCİD: 0000-0002-4250-7470
[b] ORCİD: 0009-0004-6189-1596
[c] ORCİD: 0000-0003-2487-563X

Received : 13.09.2025
Accepted : 20.01.2026

Correspondence

Ecem İpek ALTINOK
Ordu University, Faculty of Medicine, Department of Child Health and Diseases, Ordu - TÜRKİYE

ecemipekoner@gmail.com

Objective: This study aimed to evaluate and compare the performance of three widely used chatbots—ChatGPT, Gemini, and Copilot—in providing accurate and reliable answers to frequently asked questions (FAQs) related to pediatric celiac disease (CD).

Materials and Methods: A 40-item FAQ set was developed based on international guidelines and recent review articles, covering definitions, diagnosis, clinical features, laboratory tests, complications, treatment, and follow-up. Each question was independently posed in Turkish to ChatGPT, Gemini, and Copilot in August 2025 using new sessions to minimize contextual bias. Responses were blindly evaluated by a pediatric gastroenterologist, a pediatrician, and a pediatric surgeon with celiac disease. Answers were classified as: (1) comprehensive/accurate, (2) incomplete/partially accurate, (3) mixed/misleading, or (4) incorrect/irrelevant. Inter-model agreement was assessed using Cohen's kappa, and comparative statistical analyses were performed to evaluate differences in response accuracy.

Results: ChatGPT provided the highest proportion of comprehensive/accurate responses (35/40; 87.5%), followed by Gemini and Copilot (28/40; 70% each). ChatGPT demonstrated significantly higher accuracy compared with the other chatbots (χ^2 test, $p<0.05$). Copilot generated the highest rate of misleading responses (6/40; 15%). In subgroup analyses, ChatGPT performed best in treatment and follow-up questions (16/17; 94.1%), while Gemini showed relatively better performance in basic knowledge and clinical features (5/8; 62.5%) without producing misleading answers. Inter-model agreement was limited (ChatGPT–Copilot κ=0.32; Gemini–Copilot κ=0.35; ChatGPT–Gemini κ=0.11).

Conclusion: ChatGPT demonstrated the most guideline-concordant performance, whereas Copilot carried a higher risk of misleading outputs. These findings highlight both the potential and limitations of AI chatbots as first-contact tools for patient and family education, emphasizing the need for expert oversight, awareness of possible hallucinations, and guideline-based frameworks.

Key Words: *Celiac disease, pediatrics, artificial intelligence, chatbot*

Pediatrik Çölyak Hastalığı Hakkında Bilgi Sağlamada Üç Yapay Zekâ Chatbotunun Karşılaştırmalı Değerlendirilmesi

Amaç: Bu çalışmanın amacı, çocukluk çağı çölyak hastalığı (ÇH) ile ilgili sık sorulan sorulara (SSS) doğru ve güvenilir yanıt verme açısından yaygın olarak kullanılan üç sohbet motorunun—ChatGPT, Gemini ve Copilot—performanslarını değerlendirmek ve karşılaştırmaktır.

Gereç ve Yöntem: Uluslararası kılavuzlar ve güncel derleme makaleler temel alınarak; tanım, tanı, klinik bulgular, laboratuvar testleri, komplikasyonlar, tedavi ve izlem başlıklarını kapsayan 40 maddelik bir SSS seti oluşturuldu. Her soru, bağlamsal yanlılığı en aza indirmek amacıyla Ağustos 2025'te yeni oturumlar kullanılarak Türkçe olarak ChatGPT, Gemini ve Copilot'a ayrı ayrı yöneltildi. Elde edilen yanıtlar, bir çocuk gastroenteroloğu, bir pediatrist ve çölyak hastalığı bulunan bir çocuk cerrahı tarafından körleme olarak değerlendirildi. Yanıtlar; (1) kapsamlı/doğru, (2) eksik/kısmen doğru, (3) karma/yanıltıcı ve (4) yanlış/ilgisiz olmak üzere dört kategoride sınıflandırıldı. Modeller arası uyum Cohen's kappa katsayısı ile değerlendirildi ve yanıt doğruluğundaki farklar karşılaştırmalı istatistiksel analizlerle incelendi.

Bulgular: ChatGPT, kapsamlı/doğru yanıt oranı en yüksek olan sohbet motoru idi (35/40; %87,5); bunu Gemini ve Copilot izledi (her biri 28/40; %70). ChatGPT'nin doğruluk oranı diğer sohbet motorlarına kıyasla istatistiksel olarak anlamlı derecede yüksekti (χ^2 testi, $p<0,05$). Yanıltıcı yanıt oranı en yüksek olan model Copilot'tu (6/40; %15). Alt grup analizlerinde ChatGPT, tedavi ve izlem sorularında en iyi performansı gösterirken (16/17; %94,1), Gemini temel bilgi ve klinik bulgular alanında görece daha iyi performans sergiledi ve yanıltıcı yanıt üretmedi. Modeller arası uyum düşüktü (ChatGPT–Copilot κ=0,32; Gemini–Copilot κ=0,35; ChatGPT–Gemini κ=0,11).

Sonuç: ChatGPT, kılavuzlarla en uyumlu performansı sergilerken, Copilot daha yüksek yanıltıcı yanıt riski taşımaktadır. Bulgular, yapay zekâ tabanlı sohbet motorlarının hasta ve aile eğitimi için ilk temas aracı olarak potansiyelini ortaya koymakla birlikte, uzman denetimi, olası halüsinasyonların farkında olunması ve kılavuz temelli çerçevelerin gerekliliğini vurgulamaktadır.

Anahtar Kelimeler: *Çölyak hastalığı, pediatri, yapay zekâ, sohbet robotu*

Introduction

Celiac disease (CD) is a chronic, immune-mediated, and globally prevalent disorder triggered by the ingestion of gluten-containing foods, leading to immune-mediated enteropathy characterized by villous atrophy of the small intestine. It is estimated that approximately 1 in 135 individuals worldwide is affected, and a substantial proportion of patients remain undiagnosed for prolonged periods (1). Delayed or missed diagnosis may significantly impair quality of life and contribute to complications such as iron deficiency anemia, osteoporosis, infertility, and malignancies (2).

In recent years, artificial intelligence (AI)–based chatbots have emerged as widely accessible tools for obtaining rapid health-related information, both for patients and healthcare providers. However, concerns persist regarding their medical accuracy, reliability, and potential implications for patient safety, particularly when used without professional supervision. A large-scale evaluation demonstrated that four different chatbots generated unsafe or potentially harmful medical responses in 5% to 13% of cases, highlighting a non-negligible risk in clinical contexts (3).

Within the specific context of celiac disease, previous studies have yielded mixed but generally encouraging results. One comparative analysis reported that ChatGPT and similar chatbot models were able to provide clear and largely accurate responses to basic CD-related questions, although substantial variability across models and question types was observed (4). Similarly, in an evaluation of gluten-free diet planning, a weekly menu generated by ChatGPT included 75 food items, all of which were confirmed as gluten-free by a registered dietitian, demonstrating a high level of factual accuracy in dietary recommendations (5). More recently, a study from Turkey reported that ChatGPT's responses to 20 frequently asked CD-related questions were rated highly in terms of both reliability and usefulness, with Cronbach's alpha values of 0.839 and 0.753, respectively, supporting its potential role as an informational aid (6).

Despite these promising findings, important limitations remain, particularly regarding the consistency of responses, conceptual accuracy, and variability between different AI models. Moreover, most existing studies have focused on a single chatbot model, limiting the ability to draw comparative conclusions. Accordingly, there is a need for systematic, head-to-head evaluations of multiple AI-based chatbots within a standardized clinical framework.

Therefore, the present study aimed to systematically compare the accuracy and reliability of responses generated by three widely used AI-based chatbots—ChatGPT, Gemini, and Copilot— to a set of frequently asked questions related to pediatric celiac disease. We hypothesized that significant differences would be observed among the chatbots, with ChatGPT demonstrating superior overall accuracy and consistency compared with the other models.

Materials and Methods

Research and Publication Ethics: As this study did not involve human participants, patient data, or animal subjects, formal approval from an ethics committee was not required. Nevertheless, all methodological procedures were conducted in accordance with internationally accepted ethical standards for scientific publication.

This study was designed as a cross-sectional, observational comparative analysis aimed at evaluating the accuracy, consistency, and reliability of responses generated by three widely used artificial intelligence–based chatbot applications (ChatGPT, Google Gemini, and Microsoft Copilot) to questions related to pediatric celiac disease.

Development of the Question Set: The question pool was developed through a structured review of current literature and international clinical guidelines. In particular, the European Society for Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) guidelines for celiac disease, along with recent high-impact review articles, were used as the primary reference sources. Based on these materials, a total of 40 clinically relevant questions were formulated to comprehensively cover the major domains of pediatric celiac disease, including diagnosis, clinical features, laboratory findings, complications, treatment, and follow-up (7–9).

The sample size was determined using a predefined and reproducible question-based framework rather than a conventional power analysis, as the unit of analysis in chatbot evaluation studies is the question–response pair, not human participants. The selection of 40 questions was intended to ensure adequate representation of all key clinical domains while enabling balanced and standardized comparisons across the three chatbot models, consistent with the methodology of previously published AI evaluation studies (6).

Chatbot Query Procedure: Each of the 40 questions was independently submitted to all three chatbot platforms in August 2025. To minimize contextual and carry-over bias, each question was entered in a separate, newly initiated chat session using the "New Chat" function. All questions were posed in Turkish, presented in an identical format, and submitted without additional prompts, clarifications, follow-up questions, or explanatory context (Table 1).

Table 1. Forty Key Questions on Pediatric Celiac Disease

No	Question	ChatGPT	Gemini	Copilot
Basic Information				
1	What is celiac disease?	1	2	1
2	What are the most common symptoms of celiac disease in children?	1	1	1
3	Can celiac disease occur at any age?	2	2	2
4	Is celiac disease genetic?	1	1	1
5	What are the extraintestinal manifestations of celiac disease?	1	1	1
6	How does celiac disease affect growth and development in children?	1	1	1
7	Which children are at risk for celiac disease?	3	2	3
8	Which diseases are commonly associated with celiac disease?	3	2	3
Screening and Diagnosis				
9	Which blood tests are used in the diagnosis of celiac disease?	1	1	1
10	What is the anti-tTG antibody, and what does it indicate?	1	1	3
11	How is EMA (endomysial antibody) used in diagnosis?	1	1	2
12	Why should total IgA levels be measured?	1	1	1
13	How effective are genetic tests in the diagnosis of celiac disease?	1	1	2
14	Which tests are recommended for screening children for celiac disease?	1	1	1
15	When should children with a family history of celiac disease be screened?	2	1	1
16	Can false-negative test results occur in celiac disease?	1	2	1
17	Is biopsy always necessary for the diagnosis of celiac disease?	1	1	1
18	How should diagnosis be made if serological tests are negative?	1	2	1
19	What is the role of endoscopy in diagnosing celiac disease?	1	2	1
20	What are the typical biopsy findings in celiac disease?	1	1	1
21	Which diseases can be confused with celiac disease?	1	2	3
22	Is radiological imaging (e.g., barium studies) necessary in celiac disease?	1	1	1
23	What conditions can mimic celiac disease?	1	2	3
Treatment and Follow-Up				
24	What is the treatment for celiac disease?	1	1	1
25	How is a gluten-free diet implemented?	1	1	1
26	What happens if a child with celiac disease accidentally consumes gluten?	2	1	3
27	Is only wheat prohibited in the diet, or are barley and rye also forbidden?	1	2	2
28	Can oats be safely consumed by patients with celiac disease?	1	1	1
29	What should be considered regarding processed foods for children with CD?	1	1	1
30	Is it necessary to maintain a gluten-free diet for life?	1	1	1
31	What complications may occur if celiac disease is left untreated?	1	2	2
32	How soon do symptoms improve after starting a gluten-free diet?	1	1	1
33	What does refractory celiac disease mean?	1	1	1
34	Why does iron deficiency anemia occur in celiac disease?	1	2	2
35	Is there an increased risk of osteoporosis in celiac disease?	1	1	1
36	Does celiac disease increase the risk of cancer?	1	1	1
37	What is the relationship between celiac disease and type 1 diabetes in children?	1	1	1
38	How should regular follow-up be conducted in celiac disease?	1	1	1
39	Can celiac disease be completely cured?	1	1	1
40	What should be considered in the school and social life of children with CD?	1	1	1

Evaluation scale: 1 = Completely accurate and sufficient; 2 = Incomplete/partially accurate; 3 = Mixed accurate–inaccurate / misleading; 4 = Completely incorrect / irrelevant. Each question was asked in an independent session and in the same format. Responses were blindly evaluated by two specialists, and in cases of disagreement, consensus was achieved with the involvement of a third adjudicator.

Evaluation of Responses: The chatbot responses were independently and blindly evaluated by a multidisciplinary panel consisting of a pediatric gastroenterologist, a pediatrician, and a pediatric surgeon with expertise in celiac disease. Each response was assessed in comparison with current guideline-based recommendations and classified into one of four predefined categories: 1. Completely accurate and sufficient, 2. Partially accurate or incomplete, 3. Mixed/misleading or partially incorrect, 4. Completely incorrect or irrelevant.

This classification framework was designed to assess both the scientific accuracy and the clinical applicability of the chatbot-generated information.

Statistical Analysis: Statistical analysis was performed to evaluate both response distribution patterns and inter-chatbot agreement. Descriptive statistics were used to summarize the responses provided by each chatbot, expressed as counts and percentages. The proportion of "comprehensive/ accurate" responses was defined as the primary performance indicator.

To assess the level of agreement between chatbot responses, Cohen's kappa coefficient was calculated and interpreted according to standard benchmarks. Comparisons of response-category distributions among the three chatbots were conducted using the chi-square test, while Fisher's exact test was applied when expected cell counts were small. A two-sided *p* value <0.05 was considered statistically significant. All analyses were performed using IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA).

Results

Based on the analysis of 40 predefined questions, all three chatbots demonstrated overall high levels of accuracy, although notable differences in response quality were observed among the models. The highest proportion of "comprehensive/accurate" responses was generated by ChatGPT-4o (87.5%), whereas both Gemini and Copilot achieved lower accuracy rates (70.0% each). Copilot produced a comparatively higher proportion of misleading responses (15.0%), while no misleading responses were identified in Gemini's outputs (Table 2, Figure 1). Overall, the distribution of response accuracy categories differed significantly among the three chatbots (p<0.001).

In the Basic Knowledge and Clinical Features subgroup (Questions 1–8), ChatGPT and Copilot demonstrated comparable accuracy (5/8; 62.5% each), whereas Gemini showed a slightly lower accuracy rate (4/8; 50.0%). Misleading responses were observed exclusively in ChatGPT and Copilot (2/8; 25.0%), while Gemini did not generate any misleading answers (Table 3). These findings suggest that Gemini provided more consistent responses within this domain, whereas variability in text generation by ChatGPT and Copilot occasionally resulted in overlap between conceptual categories. However, the differences among the three chatbots in this subgroup were not statistically significant (p=0.21).

For questions related to Diagnosis and Testing (Questions 9–23), ChatGPT demonstrated markedly superior performance, achieving an accuracy rate of 93.3% (14/15). In contrast, both Gemini and Copilot showed lower accuracy (10/15; 66.7% each). Gemini's reduced performance was primarily attributable to a higher proportion of incomplete or partially accurate responses (3/15; 20.0%), whereas Copilot generated both incomplete (2/15; 13.3%) and misleading responses (3/15; 20.0%) (Table 4). The differences in response accuracy among the chatbots in this domain were statistically significant (p<0.001).

In the domain of Treatment, Complications, and Follow-up (Questions 24–40), ChatGPT again exhibited the highest level of accuracy (16/17; 94.1%), followed by Gemini (14/17; 82.4%, including 3/17; 17.6% incomplete responses) and Copilot (13/17; 76.5%, with 3/17; 17.6% incomplete and 1/17; 5.9% misleading responses) (Table 5). A statistically significant difference in performance was also observed among the three chatbots in this domain (p=0.03).

Analysis of inter-model agreement using Cohen's kappa coefficients revealed limited consistency across chatbot responses. Low-to-moderate agreement was observed between ChatGPT and Copilot (κ=0.32) and between Gemini and Copilot (κ=0.35), **whereas** agreement between ChatGPT and Gemini was low (κ=0.11). These findings indicate substantial variability in response categorization across models and suggest that differences in underlying information-processing and text-generation strategies may contribute to this inconsistency (Table 6).

Table 2. Distribution of chatbot responses across all accuracy categories

Response Category	ChatGPT (n=40)	Gemini (n=40)	Copilot (n=40)
Comprehensive/Accurate (1)	35 (87.5%)	28 (70.0%)	28 (70.0%)
Incomplete/Partially Accurate (2)	3 (7.5%)	12 (30.0%)	6 (15.0%)
Mixed/Misleading (3)	2 (5.0%)	0 (0.0%)	6 (15.0%)
Completely Incorrect/Irrelevant (4)	0 (0.0%)	0 (0.0%)	0 (0.0%)

* The overall distribution of response categories differed significantly among the three chatbots (chi-square test, p<0.001).

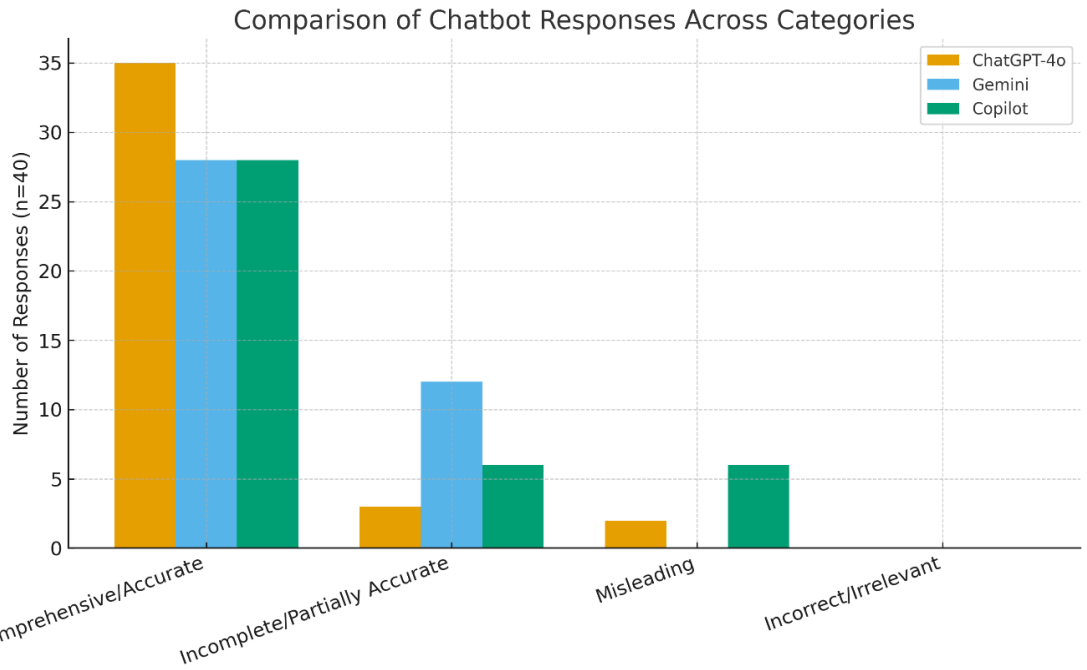

Figure 1. Comparison of chatbot responses across four accuracy categories (n = 40). ChatGPT-4o achieved the highest number of comprehensive/accurate answers (35/40; 87.5%). Gemini provided a lower proportion of comprehensive/accurate answers (28/40; 70.0%) but produced no misleading responses. Copilot demonstrated a similar overall accuracy to Gemini (28/40; 70.0%) but generated the highest proportion of misleading responses (6/40; 15.0%). These findings highlight both the strengths and limitations of each model in providing reliable information on pediatric celiac disease.

Table 3. Distribution of chatbot response accuracy by subgroup: Basic Knowledge & Clinical Features (Q1–8)

Chatbot	Comprehensive/ Accurate	Incomplete/ Partially Accurate	Mixed/Misleading	Completely Incorrect/Irrelevant
ChatGPT-4o	5 (62.5%)	1 (12.5%)	2 (25.0%)	0 (0.0%)
Gemini	4 (50.0%)	4 (50.0%)	0 (0.0%)	0 (0.0%)
Copilot	5 (62.5%)	1 (12.5%)	2 (25.0%)	0 (0.0%)

* No statistically significant difference was observed among the chatbots in this subgroup (chi-square or Fisher's exact test, p=0.21).

Table 4. Distribution of chatbot response accuracy by subgroup: Diagnosis and Testing

Chatbot	Comprehensive/ Accurate	Incomplete/ Partially Accurate	Mixed/Misleading	Completely Incorrect/Irrelevant
ChatGPT	93.3% (14)	6.7% (1)	0.0% (0)	0.0% (0)
Gemini	66.7% (10)	33.3% (5)	0.0% (0)	0.0% (0)
Copilot	66.7% (10)	13.3% (2)	20.0% (3)	0.0% (0)

* The distribution of response categories differed significantly among the three chatbots (chi-square or Fisher's exact test, p<0.001).

Table 5. Distribution of chatbot response accuracy by subgroup: Treatment, Complications, and Follow-up

Chatbot	Comprehensive/ Accurate	Incomplete/ Partially Accurate	Mixed/Misleading	Completely Incorrect/Irrelevant
ChatGPT	94.1% (16)	5.9% (1)	0.0% (0)	0.0% (0)
Gemini	82.4% (14)	17.6% (3)	0.0% (0)	0.0% (0)
Copilot	76.5% (13)	17.6% (3)	5.9% (1)	0.0% (0)

* A statistically significant difference in response-category distribution was observed among the three chatbots (chi-square or Fisher's exact test, p=0.03).

Table 6. Inter-model agreement (Cohen's kappa values) among chatbots

Comparison	Cohen's κ	Level of Agreement
ChatGPT vs. Copilot	0.32	Low–Moderate
ChatGPT vs. Gemini	0.11	Low
Gemini vs. Copilot	0.35	Low–Moderate

* Cohen's kappa values indicate low to low–moderate agreement between chatbot responses, suggesting limited consistency in response categorization across models when addressing identical clinical questions.

Discussion

In this study, when the responses of ChatGPT-4o, Gemini, and Copilot to questions on pediatric celiac disease (CD) were compared, ChatGPT demonstrated a superior profile in terms of both accuracy (comprehensive/accurate 87.5%). In contrast, Gemini and Copilot achieved lower overall accuracy rates (both 70.0%). Notably, Copilot generated a higher proportion of misleading responses (15.0%), whereas Gemini produced none. This finding is consistent with the results of a comparative study on nocturnal enuresis (10). In a 2023 study focusing on CD-related questions, expert reviewers rated the responses of ChatGPT/Bard-type models as "accurate/appropriate" in the range of 72.5–92.5%. However, variability in performance metrics and limitations in source referencing were identified (4). In the same study, inter-rater agreement was reported as moderate (κ≈0.45). By contrast, our study demonstrated only limited inter-model agreement (ChatGPT–Copilot κ=0.32; Gemini–Copilot κ=0.35; ChatGPT–Gemini κ=0.11), highlighting considerable variability in response categorization across models.

In the Basic Knowledge and Clinical Features subgroup, some responses revealed conceptual ambiguity. This was particularly evident in the failure to clearly distinguish comorbidity from complication and risk factor from symptom. For example, in response to the question "Which diseases are commonly associated with celiac disease?", Copilot grouped comorbidities (e.g., type 1 diabetes, autoimmune thyroiditis, Down syndrome) together with complications/outcomes (e.g., iron deficiency anemia, osteoporosis, growth retardation, and, rarely, intestinal lymphoma). Similarly, in the question "Which children are at risk for celiac disease?", both ChatGPT and Copilot presented risk factors (e.g., first-degree relatives, HLA-DQ2/DQ8 positivity, associated autoimmunities) alongside clinical symptoms (e.g., chronic diarrhea, abdominal pain, growth failure), creating pedagogical confusion. These conceptual distinctions should be clarified in accordance with guideline-based definitions (7–9).

In diagnostic and testing questions, Copilot produced three misleading responses (3/15; 20.0%) and two incomplete answers (13.3%), indicating that the model may sometimes make erroneous generalizations when elaborating diagnostic processes. Gemini achieved a lower accuracy rate (10/15; 66.7%) with a relatively high proportion of incomplete/partially accurate responses (5/15; 33.3%). ChatGPT demonstrated superior performance in this domain (14/15; 93.3%), with only one incomplete response (6.7%) and no misleading answers, a reassuring finding in terms of clinical information safety. At the level of specific questions, Gemini's response to "How should the diagnosis be made if serological tests are negative?" was insufficient, as it referred only to biopsy. In such cases, IgA deficiency should first be excluded, followed by the use of IgG-based tests (DGP-IgG, tTG-IgG, EMA-IgG). If serology remains negative but clinical suspicion is high, duodenal biopsy, HLA-DQ2/DQ8 positivity, exclusion of alternative enteropathies, and response to a gluten-free diet should all be considered together (seronegative celiac disease) (7–9). Likewise, in response to "Which conditions can mimic celiac disease?", ChatGPT and Copilot incorrectly listed iron deficiency anemia and type 1 diabetes—both of which are comorbidities or outcomes rather than differential diagnoses. True differentials in children should include wheat allergy, non-celiac gluten sensitivity (NCGS), Giardia infection, inflammatory bowel disease, CVID/autoimmune enteropathy, SIBO, tropical sprue, drug-induced enteropathy (e.g., olmesartan), and eosinophilic gastroenteritis. The omission of NCGS in ChatGPT's answer represented another deficiency (8, 9).

In the scenario regarding accidental gluten ingestion in children with CD, Copilot's recommendations were partially appropriate in principle but inconsistent in terms of patient safety and evidence-based practice. The statement "drinking plenty of water helps eliminate gluten from the body" is biologically unfounded; fluid intake serves rehydration purposes, and oral rehydration solution is more appropriate when vomiting or diarrhea is present. Likewise, the recommendation of probiotics has limited evidence, and due to the possibility of triggering acute lactose intolerance, products such as yogurt or kefir should be approached with caution (preferably gluten-free and lactose-free options) (8, 9). Although Copilot's response struck a partial balance between factual accuracy and practical usefulness, misleading claims such as "gluten elimination with water" and the lack of nuance regarding probiotics/dairy products could lead to misinterpretations in family education and self-care decisions. Furthermore, previous literature has reported that chatbots may occasionally generate unsafe recommendations in different clinical domains, posing potential risks to patient safety (3).

Recent literature has increasingly emphasized that, despite the growing integration of large language models into health information platforms, these systems may generate clinically inaccurate or potentially unsafe recommendations when responding to patient-posed medical questions. In a large-scale evaluation, Draelos et al. demonstrated that AI-based chatbots can provide unsafe medical advice across multiple domains, raising significant concerns regarding patient safety (11). Similarly, studies examining patient acceptance of AI-led chatbots in healthcare have highlighted the need for

transparency, reliability, and appropriate clinical oversight to ensure safe use in medical contexts (12). Taken together, these findings reinforce the importance of expert supervision and critical appraisal when chatbots are used for chronic diseases requiring long-term management and patient education, such as celiac disease.

Similarly, positive findings were observed in the area of gluten-free diet counseling. Reports presented at ACG highlighted ChatGPT's general accuracy regarding gluten-free food recommendations, while also noting limitations related to cultural preferences and source attribution (5). Our findings also demonstrate that guideline-based frameworks are essential, particularly when dealing with diagnostic nuances and conceptual classifications.

This study has several limitations. Its cross-sectional, single-time-point design means that updates in large language model (LLM) versions and knowledge bases over time may affect reproducibility and generalizability. Although two independent reviewers and one adjudicator performed blinded assessments, subjective interpretation and the inherently flexible boundaries of classification categories may have introduced bias. The study was limited to responses in Turkish, which may reduce international generalizability due to language- and culture-specific variations in expression. Finally, model settings, intra-day/session variability, and the lack of transparency regarding the generation processes are methodological factors that may affect the external validity of the findings.

In conclusion, the evaluation of responses to pediatric celiac disease questions, ChatGPT-4o demonstrated the highest accuracy (87.5%). Gemini and Copilot both achieved 70.0% accuracy; however, Copilot produced 15.0% misleading responses, whereas Gemini generated none. The low Cohen's kappa values (κ=0.11–0.35) indicate considerable variability across models. ChatGPT showed a more guideline-concordant profile, particularly in laboratory-based diagnosis and treatment/follow-up. These findings support the potential of chatbots as tools for patient and family education, while underscoring the risks of relying on them without expert supervision.

References

1. Caio G, Volta U, Sapone A, et al. Celiac disease: A comprehensive current review. BMC Med 2019; 17: 142.
2. Rondanelli M, Faliva MA, Gasparri C, et al. Micronutrients dietary supplementation advices for celiac patients on long-term gluten-free diet with good compliance: A review. Medicina (Kaunas). 2019; 55(7): 337.
3. Draelos RL, Afreen S, Blasko B, et al. Large language models provide unsafe answers to patient-posed medical questions. arXiv preprint. 2025; arXiv:2507.18905.
4. Jansson-Knodell CL, Rubio-Tapia A. S1829 Investigating accuracy and performance of artificial intelligence for celiac disease information: A comparative study. Am J Gastroenterol 2023; 118(Suppl): S1360-S1361.
5. Jansson-Knodell C, Gardinier D, Weekley K, Rubio-Tapia A. Artificial intelligence for gluten-free diet advice for celiac disease. Am J Gastroenterol 2023; 118(Suppl 10): S1361.
6. Polat YH, Cankurtaran RE. The role of artificial intelligence in celiac disease support: Analyzing ChatGPT's effectiveness for healthcare providers and patients. Anatolian Curr Med J 2025; 7: 326-330.
7. Husby S, Koletzko S, Korponay-Szabó I, et al. European society paediatric gastroenterology, hepatology and nutrition guidelines for diagnosing coeliac disease 2020. J Pediatr Gastroenterol Nutr 2020; 70(1): 141-156.
8. Catassi C, Fasano A. Celiac disease. N Engl J Med 2022; 387(24): 2259-2270.
9. Al-Toma A, Volta U, Auricchio R, et al. European society for the study of coeliac disease (ESsCD) guideline for coeliac disease and other gluten-related disorders. United Eur Gastroenterol J 2019; 7(5): 583-613.
10. Boztas AE, Ensari E. Comparative analysis of three chatbot responses on pediatric primary nocturnal enuresis. J Pediatr Urol 2025: in press.
11. Draelos RL, Afreen S, Blasko B, Brazile TL, Chase N, Patel Desai D, et al. Large language models provide unsafe answers to patient-posed medical questions. NPJ Digit Med 2024; 7: 14.
12. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbots in healthcare: A mixed-methods study. Digit Health 2019; 5: 2055207619871808.