Uludağ Üniversitesi Tıp Fakültesi Dergisi
Journal of Uludağ University Medical Faculty
52, 1780297, 2026
DOI: https://doi.org/10.32708/uutfd.1780297

ORIGINAL RESEARCH

Evaluating AI Chatbots for Pediatric Contact Lenses: A Study on Accuracy, Readability, and Reliability

Mehmet Ömer KIRIŞTIOĞLU[1], Meral YILDIZ[1], Sevde İŞLEKER[1], Esin SÖĞÜTLÜ SARI[1], Ahmet Tuncer ÖZMEN[2], Mehmet BAYKARA[1]

[1] Department of Ophthalmology, Bursa Uludağ University, Faculty of Medicine, Bursa, Türkiye.
[2] Private Practice, Bursa, Türkiye.

ABSTRACT

This study evaluated the accuracy, readability, and comprehensiveness of patient-facing responses generated by LLM-based chatbot platforms to pediatric contact lens (CL)–related questions, using expert grading and readability benchmarking. Five platforms (ChatGPT-4o, Gemini 1.5, Perplexity, Copilot, and Claude 3.5 Sonnet) were assessed using 28 curated questions. Two pediatric ophthalmologists graded anonymized outputs using DISCERN and PEMAT-P, 5-point Likert scales for accuracy and comprehensiveness, and multiple automated readability indices. Expert-written responses were included only for readability benchmarking. ChatGPT-4o produced the longest responses ($p<0.0001$). Accuracy and comprehensiveness differed across platforms ($p=0.0216$ and $p=0.0067$), with ChatGPT-4o scoring higher than Perplexity in post-hoc comparisons ($p=0.0173$ and $p=0.0087$). Expert responses were shorter but showed higher complexity on readability indices. Accuracy-based reproducibility was high for general pediatric CL queries but lower for aphakic CL–related questions ($p=0.041$), and factual inaccuracies were more frequent in aphakic topics. While LLMs may support patient education, variability in correctness and completeness underscores the need for expert oversight; these tools should complement, not replace, clinical expertise in pediatric CL usage.
Keywords: Contact Lenses. Pediatric Ophthalmology. Artificial Intelligence. Patient Education.

Pediatrik Kontakt Lensler için Büyük Dil Modeli Sohbet Botlarının Değerlendirilmesi: Doğruluk, Okunabilirlik ve Güvenilirlik

ÖZET

Bu çalışma, pediatrik kontakt lenslerle ilgili sorulara verilen yapay zekâ tabanlı sohbet robotu yanıtlarının doğruluk, okunabilirlik ve kapsamlılık açısından değerlendirilmesini, uzman değerlendirmeleri ve okunabilirlik ölçütleri kullanarak incelemiştir. ChatGPT-4o, Gemini 1.5, Perplexity, Copilot ve Claude 3.5 Sonnet olmak üzere beş büyük dil modeli, 28 adet seçilmiş soru ile test edilmiştir. Yanıtlar, DISCERN ve PEMAT-P gibi doğrulanmış araçlar, doğruluk ve kapsamlılık için 5 puanlık Likert ölçekleri ve çeşitli okunabilirlik indeksleri kullanılarak, iki pediatrik oftalmoloji uzmanı tarafından değerlendirilmiştir. Uzman yanıtları yalnızca okunabilirlik karşılaştırmalarında kullanılmıştır. ChatGPT'nin yanıtları en uzun ($p<0,0001$) ve en ayrıntılı olanlardı. Doğruluk ve kapsamlılık skorları modeller arasında anlamlı farklılık göstermiş ($p=0,0216$, $p=0,0067$) ve ChatGPT, Perplexity'den daha iyi performans sergilemiştir ($p=0,0173$, $p=0,0087$). Uzman yanıtları daha kısa olmakla birlikte okunabilirlik indekslerinde daha yüksek karmaşıklık göstermiştir. Tekrarlanabilirlik genel pediatrik kontakt lens sorularında yüksek bulunurken, afakik lenslerle ilgili sorularda anlamlı derecede düşük saptanmıştır ($p=0,041$). Özellikle afakik kontakt lens konularında bazı olgusal hatalar tespit edilmiştir. Büyük dil modelleri, hasta eğitimi materyallerini erişilebilir hale getirse de doğruluk ve bütünlükteki değişkenlik uzman gözetiminin önemini vurgulamaktadır. Bu çalışma, uzmanların sohbet robotu yanıtlarını değerlendirmesini yansıtmakta olup, uzman yanıtlarıyla doğrudan bir karşılaştırma sunmamaktadır. Yapay zekâ sohbet robotları, pediatrik oftalmolojide klinik uzmanlığın yerini almak yerine onu tamamlayıcı bir araç olabilir.
Anahtar Kelimeler: Kontakt Lens. Pediatrik Oftalmoloji. Yapay Zekâ. Hasta Eğitimi.

Date Received: 23.September.2025
Date Accepted: 26.February.2026

Dr. Mehmet Ömer KIRIŞTIOĞLU
Bursa Uludağ University,
Faculty of Medicine,
Görükle Campus, Bursa, Türkiye
E-mail: mokiristioglu@uludag.edu.tr

AUTHORS' ORCID INFORMATION
Mehmet Ömer KIRIŞTIOĞLU: 0000-0001-8010-0105
Meral YILDIZ: 0000-0002-8503-5637
Sevde İŞLEKER: 0000-0002-7352-7044
Esin SÖĞÜTLÜ SARI: 0000-0003-3729-6178
Ahmet Tuncer ÖZMEN: 0000-0002-1261-5120
Mehmet BAYKARA: 0000-0002-5555-1649

In recent years, large language models (LLMs) have become widely accessible through consumer-facing chatbots, such as ChatGPT, Copilot, and Gemini, allowing users to obtain health-related information in conversational form[1,2]. While these systems can rapidly generate patient-oriented explanations, concerns persist regarding accuracy, completeness, and the risk of confidently stated misinformation—particularly in niche and specialized clinical domains3. Pediatric contact lens (CL) care is a particularly relevant use case because safe guidance depends on nuanced counselling, age-specific risk stratification, and individualized clinical decision-making4.

The safety of CLs in children remains incompletely characterized, in part because pediatric populations are frequently excluded from clinical trials. Although preliminary evidence suggests that CL use can be safe in children, further research is needed to define a comprehensive safety profile and quantify age-specific risks[5]. In this context, accurate information dissemination and careful guidance from healthcare professionals are essential, especially as CL use among children continues to increase[6]. Globally, it is estimated that over 140 million people wear CLs, with a meaningful proportion of users being children under the age of 15[7-9]. Despite benefits for conditions such as high refractive error, aphakia after congenital cataract surgery, albinism, and nystagmus, safety concerns persist; notably, nearly a quarter of pediatric emergency room visits related to medical devices involve CLs[8,10,11].

With the increasing use of LLMs in healthcare, evaluating their ability to provide reliable and accurate information has become more intriguing and subject to scrutiny, especially in highly niche areas such as pediatric CL care. However, to date, no study has assessed the accuracy or reliability of LLM-generated responses specifically regarding pediatric CLs. This study assesses the accuracy, relevance, and comprehensiveness of information from various LLMs, comparing their outputs to expert knowledge. We aim to highlight these LLMs current capabilities and limitations in pediatric ophthalmic care and contribute to the broader discussion on integrating AI in healthcare to improve patient outcomes and safety.

Materials and Methods

Large language model search protocol

Ethics approval was not required because the study involved no human participants or identifiable data. This cross-sectional evaluation is reported in accordance with STROBE; TRIPOD-LLM items were applied where applicable[12]. The authors are not associated with or involved in the developers of the five LLMs. These five LLMs—ChatGPT-4o (OpenAI; San Francisco, CA), Gemini 1.5 (Google AI, Alphabet, CA), Perplexity (Perplexity AI, Inc.; San Francisco, CA), Copilot (Microsoft; Redmond, WA), and Claude 3.5 Sonnet (Anthropic PBC; San Francisco, CA)—were selected because they are widely used in the literature, differ in capabilities, are accessible from Türkiye, and offer free access[13-15]. We use "LLM-based chatbot" for the evaluated platforms and "LLM/model" for the underlying model; platform-level features may influence outputs. Free accounts were created for LLMs on their respective websites. Inference parameters (e.g., temperature, max tokens, and random seed) were not user-accessible in the free web interfaces and therefore could not be standardized across platforms. The questions were entered into the LLMs on the same day (September 14, 2024). A detailed summary of model access conditions and available settings for each LLM platform as of the evaluation date (September 14, 2024) is provided in the Supplementary Material 1 to ensure transparency and reproducibility. The conversation history was deleted before and after each query to ensure independent responses. Intended audience and prompting framework: All prompts were designed to simulate questions asked by parents/guardians seeking patient-facing information about pediatric CL wear and care, rather than clinician-to-clinician decision support. To standardize tone across models, each new session began with the same introductory statement ("I have some questions about CLs in children" or "I have some questions about aphakic CLs in babies and infants"), followed by the specific question. Then, the specific questions were asked. Only free-access versions of the LLMs were evaluated, which may have limitations in terms of response depth, access to real-time data, and potential biases due to training constraints or restricted datasets.

Each question was phrased to reflect the perspective of a parent/guardian seeking patient-facing guidance on pediatric CL wear and care. The initial question pool was compiled from publicly available patient-education webpages that commonly appear in first-page search results for "pediatric contact lenses," and items were rephrased into concise FAQ-style prompts; brief follow-up prompts were added only when necessary to resolve ambiguity. An expert panel (n = 2) of board-certified pediatric ophthalmologists reviewed the questions for plausible caregiver wording and topic relevance. Redundant items were removed, and the remaining questions were revised to standardize scope and phrasing. The panel further suggested additional questions to broaden coverage across hygiene and handling, wear schedules, infection prevention, activity-related risks (e.g., swimming), follow-up recommendations, and red-flag symptoms, including infant aphakic CL management where applicable. The research team integrated these recommendations to finalize the 28-question FAQ set used to prompt the LLMs. The final set of 28 questions was curated from the sources in the Supplementary Material 2.

The questions were adapted to focus on pediatric aphakic CLs where necessary. The following types of inquiries were excluded from the study: questions with similar meanings, subjective or highly specialized questions (e.g., "Are daily contact lenses safe for children with seasonal allergies? "), vague inquiries (e.g., "Why do my child's eyes feel sticky after wearing lenses? "), and non-medical questions related to CL use (e.g., "What is the best brand of lenses for children? "). Additionally, questions aimed at improving understanding of pediatric CL safety and care (e.g., "How do contact lenses work? ") and complex questions were included to assess the model's capacity to assist parents and patients (e.g., "What are the risks of using lenses in children with keratoconus? "). One author conducted the exclusion decision process, and the second author subsequently validated and approved the selected inquiries. A total of 28 questions specific to pediatric and aphakic CLs were posed to the LLMs (Table I). The outputs from each LLM were collected and compared in a blind fashion. LLMs responses were exported in plain text format to prevent bias by a research assistant unaffiliated with the grading process. These responses were anonymized and uploaded to a third-party survey platform (SurveyMonkey, San Mateo, CA, USA) and a Microsoft Excel (Microsoft, Redmond, WA, USA) sheet to ensure an unbiased assessment and minimize order bias. For randomization, the five LLMs responses for each question were shuffled into a single randomized order, and this same order was presented to both graders. In cases where the ophthalmologists' accuracy or comprehensiveness ratings differed, the discrepancies were discussed in a consensus meeting, and the agreed score was used for the final analysis.

Assessment of reproducibility

Each of the 28 questions was submitted twice to each LLM-based platform on the same day (September 14, 2024) in independent reset sessions, with conversation history cleared before and after each query. Two blinded pediatric ophthalmologists graded both outputs using the same 5-point Likert accuracy scale described above. Overall accuracy for each model was summarized across both runs (e.g., mean of A1 and A2), and no tie-break rule privileging Response 1 was used. Accuracy-based reproducibility was defined as $|A1-A2| \leq 1$ and summarized as Reproducibility (%) = $100 \times (1 - |A1-A2|/4)$. Here, A1 and A2 range from 1 to 5; therefore, the maximum possible difference is 4. We also quantified response-level reproducibility by comparing Response 1 vs Response 2 using normalized Levenshtein similarity and TF–IDF cosine similarity, averaged across the 28 questions for each model. Text-similarity metrics were scaled to 0–1 and reported as percentages. Any grading discrepancies were resolved by consensus. Accuracy-based reproducibility reflects stability of correctness, whereas text-similarity metrics capture stability of wording and may differ even when accuracy is similar. This metric is independent of whether either response is correct; it captures the stability of accuracy scores across repeated generations.

Table I. The questions asked of LLMs are shown in the table and categorized by topic.

Pediatric Contact Lenses		Aphakic Contact Lenses	
No	**Question**	**No**	**Question**
Q1	*Is the risk of infections higher in children when using contact lenses?*	Q18	*Is the risk of infections higher in babies when using contact lenses?*
Q2	*What do I look for to determine whether my child has a problem with the contact lenses?*	Q19	*What do I look for to determine whether my child has a problem with the contact lenses?*
Q3	*Are contact lenses safe for kids?*	Q20	*Are aphakic contact lenses safe for babies?*
Q4	*At what age are contacts safe for kids to wear?*	Q21	*Which is better–glasses or contact lenses for my baby?*
Q5	*How to reduce the risk of contact lens complications in a child?*	Q22	*Does my baby have to wear a contact lens all of the time, even when they sleep?*
Q6	*What will happen if proper contact lens hygiene is ignored in my child?*	Q23	*How do you take care of aphakic contact lenses?*
Q7	*Are contact lenses more expensive than glasses in the pediatric age group?*	Q24	*What should I do if my baby is having a problem with their contact lens?*
Q8	*How do you take care of contact lenses?*	Q25	*How long after congenital cataract surgery my baby use aphakic contact lenses?*
Q9	*What should I do if my child is having a problem with their contact lens?*	Q26	*Which contact lens material is better for my baby: soft, silicon hydrogel, or rigid gas permeable?*
Q10	*Which is better–glasses or contact lenses for my child?*	Q27	*How do you take care of aphakic contact lenses?*
Q11	*Can the contact lens tear whilst it is in my child's eye?*	Q28	*My baby was operated on in one eye for a congenital cataract. Should my baby wear glasses in addition to the contact lens for aphakia in one eye?*
Q12	*What should I do if my child's contact lens falls out*		
Q13	*Does my child have to wear a contact lens all of the time, even when they sleep?*		
Q14	*When my child is using contact lenses what about bath times?*		
Q15	*When using contact lenses, can my child play in a sand pit / go to the beach?*		
Q16	*Can a contact lens get lost and move to the back of the eye?*		
Q17	*When using contact lenses can my child go swimming?*		

Assessment of readability

For the readability assessment, references in LLMs responses were first removed. The readability of the outputs was assessed using the Flesch Reading Ease (FRE) and Flesch–Kincaid Reading Grade Level (FKRGL) formulae. Additionally, readability indices, such as the Simplified Measure of Gobbledygook (SMOG) Index, Automated Readability Index (ARI), Linsear, Gunning Fog, and Coleman-Liau, were also evaluated as in the literature (Table II)[16]. These assessments evaluate the ease of understanding text based on syllables, words, and sentences. Readability assessments also included responses provided by two ophthalmologists experienced in pediatric CLs. The average of these two experts' scores was calculated and compared with the LLMs responses. For readability benchmarking, two pediatric ophthalmologists drafted brief patient-facing responses to the same questions; these expert-written texts were analyzed only for readability comparisons. They were not used as a gold standard for accuracy or comprehensiveness scoring.

Table II. Comparison of Readability Formulas: Descriptions and Measurement Criteria

Readability Formula	Description	Measurement Criteria
FRE (Flesch Reading Ease)	Measures text readability on a scale from 1 to 100. Higher scores indicate more readable text. A score between 70-80 approximates a 7th-grade reading level.	Scale of 1 to 100; higher score = easier to read
FKGL (Flesch-Kincaid Grade Level)	Assigns a U.S. school grade level to a text, indicating the educational level required to comprehend it.	Grade level score; e.g., 8 = 8th-grade reading level
Gunning Fog Index	Estimates the years of formal education needed to understand a text on first reading, based on sentence length and the percentage of complex words.	Education years required; considers sentence length and complex words
SMOG Index	Estimates the number of years of education required to understand a text by analyzing the number of polysyllabic words.	Education years required; based on polysyllabic words
Automated Readability Index (ARI)	Calculates the grade level of a text based on the average number of characters per word and the average number of words per sentence.	Grade level; average characters per word and words per sentence
Coleman-Liau Index	Estimates the readability of a text by analyzing the average number of characters per word and the number of sentences per 100 words.	Grade level; average characters per word and sentences per 100 words
Linsear Write Formula	Determines the readability of a text by evaluating the frequency of simple and complex words and estimating the grade level required for comprehension.	Grade level; frequency of simple and complex words

Assessment of quality of information

Each response was evaluated for information quality using two validated tools: DISCERN (with overall scores ranging from 1 [low] to 5 [high] for quality) and the Patient Education Materials Assessment Tool for Printed Materials (PEMAT-P). DISCERN was developed as a standardized tool and is often used to assess the trustworthiness and comprehensiveness of patient information materials, including websites, pamphlets, and other educational resources. The DISCERN tool provides a set of questions that guide the evaluation of health information, with each question typically scored on a scale from 1 (low) to 5 (high). The overall DISCERN score reflects the quality of the material, with higher scores indicating more reliable, well-balanced, and presented information [17]. DISCERN consists of 16 items scored 1–5, yielding a total score range of 16–80; higher scores indicate better quality. The PEMAT-P consists of 24 questions, scored as '0' for disagree, '1' for agree, or 'NA' for not applicable. The scores were summed to produce a percentage score, with higher percentages indicating better understandability and actionability. NA items were excluded from the denominator when computing PEMAT-P percentage scores. Responses were considered understandable and actionable if they scored 70% or higher based on prior validation [18]. The average score of two reviewers was calculated for each LLM response. PEMAT-P scores emphasize the understandability and actionability of patient education materials, while DISCERN scores focus on the reliability and quality of information, particularly regarding treatment options.

Assessment of accuracy and comprehensiveness

The accuracy of the LLMs responses was evaluated by two pediatric ophthalmologists with high proficiency in English. Accuracy was rated using a 5-point Likert scale: (1) Very inaccurate (contains major factual errors), (2) Somewhat inaccurate (partially correct but contains misleading information), (3) Moderately accurate (mostly correct but lacking important details), (4) Accurate (factually correct with minor omissions), (5) Very accurate (comprehensive and error-free). Comprehensiveness was scored similarly: (1) Very incomplete (major gaps in information), (2) Somewhat incomplete (important details missing), (3) Moderately comprehensive (adequate but not detailed), (4) Comprehensive (detailed with minor gaps), (5) Very comprehensive (fully detailed and informative). The graders, who were unaware of the LLMs source, rated the responses using a 5-point Likert scale. The order of responses was randomized within SurveyMonkey (San Mateo, CA, USA) to mitigate order bias. Each grader independently scored the responses before the final scores were averaged to determine accuracy and comprehensiveness ratings. The average score from all reviewers determined the total accuracy score for each LLM-generated response.

Data analysis

To assess inter-rater agreement, we calculated weighted Cohen's kappa (κ) between the two graders for Likert-scale ratings. Descriptive analyses were performed in Microsoft Excel. Repeated-measures statistical analyses were conducted in SPSS (IBM SPSS Statistics for Windows, Version 28.0; IBM Corp., Armonk, NY, USA), and figures were created using GraphPad Prism (version 10.0.0; GraphPad Software, Boston, MA, USA). Because each question was answered by all models (question-level clustering), between-model comparisons were performed at the question level using Friedman tests, followed by Holm-adjusted pairwise Wilcoxon signed-rank tests for post-hoc comparisons. The same repeated-measures approach was applied to readability indices (e.g., word/character counts, FRE, FKGL, SMOG, ARI, Linsear, Gunning Fog, Coleman–Liau) and to Likert-based outcomes (accuracy and comprehensiveness). Inter-rater agreement was calculated for Likert-scale accuracy/comprehensiveness ratings; formal inter-rater reliability was not computed for DISCERN/PEMAT. Kappa values were interpreted as <0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and >0.81 almost perfect agreement. Two-tailed p-values <0.05 were considered statistically significant. Continuous variables are reported as median (min–max) for repeated-measures analyses and as mean ± SD where appropriate.

Results

Readability differed across response sources (five LLMs and the expert mean) when analyzed at the question level using repeated-measures statistics (Friedman tests; Table III). Output length showed the strongest separation: ChatGPT-4o produced longer responses than Perplexity, Claude 3.5 Sonnet, Gemini, and Copilot in both word count and character count (Friedman p<0.001; Holm-adjusted paired tests p < 0.01 for each comparison), while the expert mean was substantially shorter than all LLM outputs (Holm p<0.001). Flesch Reading Ease (FRE) showed an overall group effect (Friedman p<0.001), but after Holm correction the main differences were driven by the expert mean versus the LLM outputs rather than by consistent differences among the LLMs themselves. In contrast, grade-level indices demonstrated model-dependent variation. Gemini tended to generate more complex text, with higher ARI and Flesch–Kincaid grade levels and higher SMOG and Gunning Fog scores compared with ChatGPT-4o and Copilot (Holm p<0.05 for significant contrasts). Coleman–Liau values also differed overall (Friedman p<0.05), with Copilot yielding lower Coleman–Liau scores than Claude and Gemini (Holm p<0.05). Linsear Write scores differed across sources (Friedman p<0.05) and were lowest for ChatGPT-4o, suggesting simpler sentence structure on this metric compared with the other LLMs and the expert mean (Holm p<0.05).

DISCERN scores revealed that the mean values (±SD) for ChatGPT, Perplexity, Claude, Gemini, and Copilot were 57.66±5.74, 52.39±8.89, 55.27±6.33, 56.04±8.40, and 54.61±11.28, respectively, with no statistically significant difference observed between the groups (p=0.2008) (Figure 1E). For PEMAT-P understandability scores, the mean values for ChatGPT, Perplexity, Claude, Gemini, and Copilot

Table III. Readability scores of responses by LLMs and expert mean.

	ChatGPT 4o	Perplexity	Claude 3.5 Sonnet	Gemini	Copilot	Expert Mean	p-value
ARI Index	9.458^{c} (6.109-15.36)	10.91^{b} (5.288-17.41)	9.542 (5.873-19.30)	11.51^{ac} (4.885-19.26)	9.154^{ab} (5.171-12.87)	11.1 (5.46–17.1)	**<0.0001** **^{abc} <0.05**
Character Count	1933^{afghi} (639-3145)	1366^{bg} (321-2396)	1292^{cf} (1083-1980)	1330^{dh} (457-2389)	1340^{ei} (509-2092)	343^{abcde} (226–809)	**<0.0001** **^{abcdefghi}<0.001**
Coleman Liau	11.58 (8.32-15.28)	12.51 (7.219-15.68)	12.26^{a} (6.138-15.59)	11.98^{b} (6.60-16.17)	11.38^{ab} (6.83-13.92)	13.1 (6.67–17.9)	**<0.0078** **^{ab}<0.05**
Flesch – Kincaid Grade Level	10.83^{ad} (8.326-16.76)	11.92 (7.59-18.47)	11.53 (7.48-18.16)	12.36^{cd} (6.72-18.62)	11.09^{bc} (6.825-13.75)	11.4^{ab} (7.49–16.1)	**<0.0001** **^{abc}<0.05**
Flesch Reading Ease	41.25^{a} (13.36 – 60.37)	38.74 (7.184-63.58)	37.55 (20.04-71.29)	38.11 (13-70.24)	41.21^{b} (27.66-68.93)	38.6^{ab} (13.9–66.8)	**<0.001** **^{ab}<0.01**
Gunning Fog	13.92^{ac} (11.32 – 20.12)	14.17 (9.616-23.6)	14.78 (9.171-20.79)	15.27^{c} (10.31-21.19)	14.01^{b} (10.12-16.21)	14.9^{ab} (10.9–20.1)	**<0.001** **^{abc}<0.05**
Linsear Write	2.730^{a} (0.6579 – 15.80)	4.638^{b} (1.783 – 21.0)	4.187^{c} (2.048 – 10.50)	4.955^{d} (1.81 – 14.80)	4.01^{e} (2.1-13.8)	3.85^{abcde} (26–62.8)	**<0.0001** **^{abcde}<0.0001**
SMOG Index	12.69^{ac} (11.04-17.35)	13.12 (9.978-19.78)	13.26 (9.467-17.63)	13.85^{c} (9.995-17.41)	12.96^{b} (10.32 – 14.70)	13.5^{ab} (10.4–17)	**< 0.001** **^{abc}<0.05**
Word Count	303^{aefgh} (108-494)	208.5^{bg} (50-355)	209.5^{cf} (178-309)	208.5^{de} (74-337)	228^{eh} (84-337)	57.2^{abcde} (36–124)	**<0.0001** **^{abcdefgh} <0.01**

Values are presented as median (min–max) across the 28 questions. Friedman test was used for overall comparisons with Holm-adjusted pairwise Wilcoxon signed-rank tests. Superscripts denote post-hoc groupings (shared letters = not significantly different).

were 82.38±6.32, 75.95±11.11, 78.75±6.26, 81.07±6.4, and 76.59±11.55, respectively, and a statistically significant difference was found between the groups (p=0.0124). Pairwise comparisons indicated that this difference was significantly higher for ChatGPT than Perplexity (p=0.0054) (Figure 1A). As for PEMAT-P actionability scores, the mean values for ChatGPT, Perplexity, Claude, Gemini, and Copilot were 76.07±6.85, 67.14±14.87, 70.95±8.94, 75.77±9.29, and 67.98±14.18, respectively, with a statistically significant difference between the groups (p=0.0005). This difference was statistically significant between three pairwise groups (ChatGPT-Perplexity: p=0.037, Perplexity-Gemini: p=0.045, and Gemini-Copilot: p=0.0432) (Figure 1B).

In terms of reproducibility, when looking at the average values of 28 questions, ChatGPT scored 94.29±6.15, Perplexity 94.02±6.02, Claude 92.95±6.2, Gemini 94.11±6.24, and Copilot 92.41±6.5. No statistically significant difference was observed between the LLMs (p=0.143) (Figure 1C). Response reproducibility was evaluated; ChatGPT showed the highest stability (Levenshtein 92.8%, cosine 94.5%), followed by Claude 3 Sonnet (88.4%, 85.2%) and Gemini (80.1%, 83.7%). Perplexity and Copilot exhibited greater between-run variability, with lower Levenshtein similarity (72.4% and 73.4%) despite higher cosine similarity (81.7% and 82.2%), suggesting that repeated queries may yield notably different wording while preserving overall content. However, when reproducibility was analyzed based on question categories, the average score for pediatric questions was 94.41±6.13, while for aphakic questions, it was 92.23±6.08, and the difference between them was statistically significant (p=0.041) (Figure 1D).

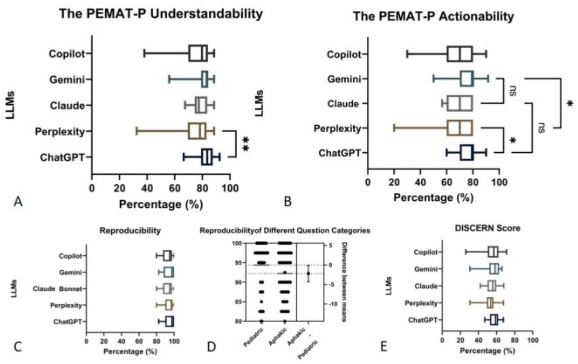

Figure 1.
*DISCERN, Reproducibility, PEMAT-P understandability, and actionability scores of the responses by the LLMs. The statistical significance is denoted as follows: *p<0.05, **p<0.01, ****p<0.001.

Based on the Likert-scale results, the median (range) accuracy scores for ChatGPT, Perplexity, Claude, Gemini, and Copilot were 4.333 (3.667–4.667), 4.000 (2.667–4.667), 4.333 (3.000–4.667), 4.333 (2.667–4.667), and 4.333 (3.333–4.667), respectively. The median values for comprehensiveness were 4.33 (range: 2.67-5), 4.0 (range: 2-4.67), 4.0 (range: 3.0-4.67), 4.33 (range: 2.0-5.0), and 4.17 (range: 2.33-5.0), respectively (Figure 2). There was a significant difference between the groups in terms of both accuracy and comprehensiveness (p=0.0216, p=0.0067). In Holm-adjusted post-hoc pairwise Wilcoxon signed-rank tests, ChatGPT-4o was the only model that differed significantly from Perplexity, with higher accuracy and comprehensiveness scores (p=0.0173 and p=0.0087, respectively; Figure 2).

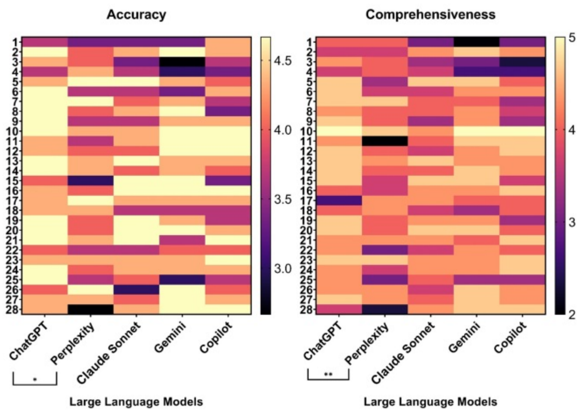

Figure 2.
Accuracy and comprehensiveness scores of the different LLMs. The key differences between LLM performance have been highlighted using annotations to emphasize statistically significant contrasts.

The inter-rater agreement between the two observers was assessed using Cohen's kappa. The weighted Cohen's kappa was calculated to account for the agreement corrected by chance. The kappa value was 0.623, indicating a substantial level of agreement between the two observers on the 5-point scale.

All LLMs examined clearly emphasized in all responses that they are not medical professionals and advised users to consult an eye care specialist for accurate diagnosis and appropriate treatment.

Discussion and Conclusion

Recent advancements in AI have transformed access to health information, allowing patients and healthcare professionals to receive instant responses and continuous educational updates[19]. Large language models (LLMs), trained on vast datasets, generate humanlike responses using deep-learning algorithms, but their effectiveness varies based on underlying architectures[20,21]. However, a key limitation is their lack of specialized training for pediatric applications, raising concerns about accuracy and reliability in niche medical fields[22].

Current models are typically trained on general medical corpora, which may not adequately address the unique nuances of pediatric healthcare, potentially leading to suboptimal patient education outcomes[22,23]. A fundamental distinction necessitates early emphasis regarding the interpretation of our findings: the structural difference between web-enabled models (Copilot, Gemini 1.5 and specifically Perplexity) and static, pre-trained models (Claude 3.5 Sonnet and ChatGPT-4o). While static models rely on frozen training datasets, web-enabled models retrieve real-time data from the internet. In the absence of universally standardized guidelines for pediatric CL care, the open web often contains conflicting information. Consequently, web-enabled models risk propagating online inaccuracies, whereas static models may offer greater stability by relying on synthesized training data rather than unverified real-time search results[24]. Rossettini et al. observed better performance with a retrieval-enabled platform (Perplexity) in lumbosacral pain management; this may be partly attributable to the ready availability of guideline-based resources online in that domain. This discrepancy likely reflects the availability of online evidence: while Rossettini et al.'s domain benefits from accessible Clinical Practice Guidelines ideal for retrieval, pediatric CL care lacks unified standards[25]. Consequently, in our guideline-naive context, static models relying on generalized reasoning proved more robust than web-enabled LLMs prone to retrieving conflicting online data. This suggests that the clinical utility of web-browsing is context-dependent: advantageous for guideline-rich conditions but potentially detrimental for niche topics lacking standardized consensus.

LLMs can improve patient education by making health information accessible and engaging, providing instant answers to common pediatric questions. The communication needs in pediatrics are unique, requiring models to generate responses that are not only medically accurate but also age-appropriate and understandable for children and their guardians. General medical LLMs may struggle with this aspect, leading to less effective patient education materials for younger patients[23,26]. In pediatric applications, there is a greater emphasis on safety, privacy, and ensuring that the information provided is suitable for children—factors that may not be as prominent in general medical contexts.

According to DISCERN and Likert scores, the 5 AI LLMs in this analysis generated high-quality and respectively accurate responses about pediatric CLs for the patients and parents and did not appear to spread misinformation. Yet, there were differences between LLMs. In the PEMAT-P, ChatGPT and Gemini outperformed other LLMs by providing higher-quality content, while Perplexity lagged. However, this difference in quality was not reflected in the DISCERN scoring. This indicates that while there were no significant differences among the LLMs regarding reliability and information quality in their responses, there were differences in understandability and actionability. ChatGPT tends to provide longer responses, which may have contributed to this difference.

There were disagreements among LLMs on some questions. For example, while Copilot and Perplexity responded affirmatively to Q15, Claude and Gemini generally answered that they could not participate. ChatGPT, on the other hand, gave a more diplomatic answer, stating that they could participate under good supervision. These differences could lead to different guidance for families.

Although rare, LLMs sometimes provide responses that could be misleading. For example, Claude incorrectly answered Q22 in Table IV. However, when re-queried, the Claude 3.5 Sonnet did not repeat this mistake, indicating low reproducibility for this question. Such incorrect answers can be tagged as AI hallucinations and could provide misleading information for parents seeking guidance[27].

Table IV. Example of AI hallucination from the Claude 3.5 Sonnet.

Question	Answer
I have some questions about aphakic contact lenses in babies and infants. Does my baby have to wear a contact lens all of the time, even when they sleep?	*"Yes, for infants and young children who have had a cataract removed, it is generally recommended that they wear their aphakic contact lenses at all times, including while sleeping.* *There are a few important reasons for this:* *1. Visual development - Proper visual input and focus through the aphakic lens is crucial for normal visual development in the infant's brain during this critical period.* *2. Eye health - Without a lens in the eye, the eye is more prone to infections, irritation, and potential complications. The contact lens protects the eye.* *3. Induced refractive error - Without the lens in place, the eye will be extremely farsighted (hyperopic), impacting vision.* *Removing the lenses during sleep interrupts the visual input and could impair the infant's visual development. It also risks the potential for the eye to develop complications when the lens is out.* *The lenses are designed for extended/continuous wear, and the eye doctor will provide specific instructions on lens removal, cleaning, and replacement schedules based on the needs of the individual child. Parental supervision is required to ensure the lenses remain properly in place."*

Readability analyses showed an overall difference across response sources (five LLMs and the expert mean) at the question level (Friedman tests; Table III). While output length separated the groups most clearly—ChatGPT-4o producing longer responses than all other LLMs, and the expert mean being substantially shorter than all LLM outputs—readability differences were more nuanced. FRE demonstrated an overall group effect, but after Holm correction the primary contrasts were driven by the expert mean versus LLM outputs rather than consistent differences among the LLMs themselves. In contrast, several grade-level indices showed model-dependent variation: Gemini tended to generate more complex text than ChatGPT-4o and Copilot (higher ARI, FKGL, SMOG, and Gunning Fog in significant contrasts), whereas Copilot yielded lower Coleman–Liau values than Claude and Gemini. Linsear Write also differed across sources and was lowest for ChatGPT-4o, highlighting that readability formulas can diverge because they emphasize different surface features (e.g., sentence length vs. word complexity), and therefore should be interpreted as complementary rather than interchangeable. Importantly, prompts were designed to simulate parent/guardian questions, and the resulting reading levels may still be demanding for many families[29]. We agree with the reviewer that interpreting these outputs as if the questions were asked by children would be inappropriate: LLMs can adapt tone and complexity to perceived user characteristics, and child-phrased prompts might elicit simpler, more age-appropriate responses. Therefore, our findings should be interpreted specifically within a parent/guardian information-seeking scenario, not as a proxy for child-directed patient education.

LLM responses were generally easier to understand than expert responses, with ChatGPT and Perplexity producing significantly higher FRE scores, indicating greater readability. The Linsear Write Index was significantly lower for all LLMs, suggesting simpler sentence structures. ChatGPT and Copilot had significantly lower ARI scores, while the Coleman-Liau Index was also significantly lower across all LLMs compared to expert responses, particularly for Copilot, Gemini, Claude, and Perplexity. These findings indicate that LLM-generated responses were written at a lower reading grade level and were easier to comprehend than those provided by ophthalmologists.

While LLMs can provide helpful information across various medical fields, their reproducibility can vary significantly based on the complexity of the questions and the specific specialty[30,31]. This may explain why, in our study, LLMs showed high reproducibility for general pediatric CL questions but lower reproducibility for the more specialized topic of aphakic CLs.

ChatGPT, currently one of the most popular LLMs, relies heavily on the quality of the data it has been trained on and typically cannot access the internet independently[32]. In contrast, other LLMs like Perplexity AI use a combination of static training and real-time information from the web. However, the reliability of online information is not always guaranteed, and it may not meet academic or peer-reviewed standards. As a result, AI LLMs' limitations stem from the datasets they are trained on and the quality of the online sources they access.

This study has several limitations that should be noted. First, the versions of the LLMs used in this evaluation may differ from future iterations, which could impact the generalizability of the findings as these models evolve. Second, the LLMs did not address specific CL subtypes, such as rigid or orthokeratology, which are essential aspects of pediatric eye care. This limitation restricts the relevance of the LLM responses to more generalized information, potentially missing critical details necessary for comprehensive care. Third, since the questions were simulated as asked by parents, the study may not fully represent how healthcare professionals or older children might engage with these AI tools, possibly influencing the depth and specificity of the responses provided. Fourth, because each prompt was sampled only twice per model, our reproducibility estimates may not fully capture within-model variability across repeated generations; studies with more runs per prompt could provide more precise stability estimates. In addition, accuracy-based reproducibility based on $|A1-A2|$ captures the magnitude of variability but does not indicate the direction of change (i.e., improvement vs deterioration) without stratified analyses. Fifth, inter-rater reliability was not formally quantified for DISCERN and PEMAT ratings (e.g., kappa/ICC); although discrepancies were resolved by consensus, the degree of baseline agreement between raters cannot be reported. Sixth, the assessment of accuracy and comprehensiveness relied solely on the consensus of two expert graders without a formal gold standard, which introduces the possibility of grader bias. Although the weighted kappa indicated substantial agreement, the lack of an objective reference standard limits reproducibility and generalizability. Another important limitation relates to the unequal baseline functionality across LLMs. At the time of data collection, Perplexity was inherently web-enabled, whereas ChatGPT-4o and Claude 3.5 Sonnet free versions lacked real-time search capability. This capability mismatch limits direct comparability across platforms; web-enabled retrieval may improve performance in guideline-rich topics but may also introduce heterogeneity or errors in niche domains. While briefly acknowledged earlier, further stratification between web-enabled and static models is warranted, and future studies should re-evaluate LLM performance using updated versions, many of which now include web search features. Finally, only the free-access versions of LLMs were analyzed. These versions may differ from premium or enterprise models in terms of dataset access, response depth, and real-time information retrieval, and biases inherent in publicly available models could have affected the accuracy and comprehensiveness of responses.

LLMs can be valuable adjuncts in pediatric CL care, but the observed variability in accuracy, comprehensiveness, readability, and between-run consistency underscores the need for cautious, supervised use and further refinement. These tools may assist clinicians in drafting patient-facing explanations, yet outputs should be verified before clinical use, and parents should be reminded that LLMs are not diagnostic devices; while they may reinforce general hygiene concepts, guidance regarding symptoms, complications, infant/aphakic care, overnight wear, or water exposure should be treated as preliminary and confirmed with an eye-care professional. Developers should prioritize pediatric-specific, high-quality training data and greater transparency about data sources, and future research should evaluate paid versions and web-enabled/real-time retrieval settings to better define safe, context-appropriate clinical utility. Overall, LLMs may improve access to information, but they should complement—not replace—professional expertise, with patient safety as the central priority.

Researcher Contribution Statement:
Idea and design: M.Ö.K., M.B.; Data collection and processing: M.Ö.K., M.Y., S.I., A.T.O.; Analysis and interpretation of data: M.Ö.K., M.Y., E.S.S.; Writing of significant parts of the article: M.Ö.K., M.Y., S.I.; Critical revision: E.S.S., M.B.
Support and Acknowledgement Statement:
Not applicable.
Conflict of Interest Statement:
The authors of the article have no conflict of interest declarations.
Ethics Committee Approval Information:
Ethics approval was not required or pursued because the study involved no patient participation.

References

1. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. Jun 3 2021;4(1):93. doi:10.1038/s41746-021-00464-x
2. Wang L, Wan Z, Ni C, et al. A Systematic Review of ChatGPT and Other Conversational Large Language Models in Healthcare. *medRxiv*. Apr 27 2024;doi:10.1101/2024.04.26.24306390
3. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*. 2023/09/22 2023;23(1):689. doi:10.1186/s12909-023-04698-z
4. Sengor T, Gencaga Atakan T. Management of Contact Lenses and Visual Development in Pediatric Aphakia. *Turk J Ophthalmol*. Apr 19 2024;54(2):90-102. doi:10.4274/tjo.galenos.2023.56252
5. Tomiyama ES, Kobia-Acquah E, Ansari SM, et al. Scoping review: Reporting characteristics for the safety of contact lenses in the pediatric population. *Optom Vis Sci*. Jul 16 2024;doi:10.1097/OPX.0000000000002156
6. Bullimore MA, Richdale K. Incidence of Corneal Adverse Events in Children Wearing Soft Contact Lenses. *Eye Contact Lens*. May 1 2023;49(5):204-211. doi:10.1097/ICL.0000000000000976
7. Ezinne NE, Bhattarai D, Ekemiri KK, et al. Demographic profiles of contact lens wearers and their association with lens wear characteristics in Trinidad and Tobago: A retrospective study. *PLoS One*. 2022;17(7):e0264659. doi:10.1371/journal.pone.0264659
8. Bullimore MA. The Safety of Soft Contact Lenses in Children. *Optom Vis Sci*. Jun 2017;94(6):638-646. doi:10.1097/OPX.0000000000001078
9. Lazarus DR. Can Children Wear Contact Lenses? Accessed 12/08/2024, 2024. https://www.optometrists.org/childrens-vision/guide-to-childrens-eye-exams/can-kids-wear-contact-lenses/
10. de Brabander J, Kok JH, Nuijts RM, Wenniger-Prick LJ. A practical approach to and long-term results of fitting silicone contact lenses in aphakic children after congenital cataract. *CLAO J*. Jan 2002;28(1):31-5.
11. Vincent SJ. The use of contact lenses in low vision rehabilitation: optical and therapeutic applications. *Clin Exp Optom*. Sep 2017;100(5):513-521. doi:10.1111/cxo.12562
12. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. Jan 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5
13. Garcia-Porta N, Vaughan M, Rendo-Gonzalez S, et al. Are artificial intelligence chatbots a reliable source of information about contact lenses? *Cont Lens Anterior Eye*. Apr 2024;47(2):102130. doi:10.1016/j.clae.2024.102130
14. Nield D. Battle of the AI bots: Copilot vs ChatGPT vs Gemini. Popular Science. 2024. https://www.popsci.com/technology/copilot-vs-chatgpt-vs-gemini/
15. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. *J Med Internet Res*. Dec 28 2023;25:e51580. doi:10.2196/51580
16. Gencer A. Readability analysis of ChatGPT's responses on lung cancer. *Sci Rep*. Jul 26 2024;14(1):17234. doi:10.1038/s41598-024-67293-2
17. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health*. Feb 1999;53(2):105-11. doi:10.1136/jech.53.2.105
18. Vishnevetsky J, Walters CB, Tan KS. Interrater reliability of the Patient Education Materials Assessment Tool (PEMAT). *Patient Educ Couns*. Mar 2018;101(3):490-496. doi:10.1016/j.pec.2017.09.003
19. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J*. Jul 2021;8(2):e188-e194. doi:10.7861/fhj.2021-0095

20. Hernandez A, Amigo JM. Attention Mechanisms and Their Applications to Complex Systems. *Entropy (Basel)*. Feb 26 2021;23(3)doi:10.3390/e23030283
21. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. Oct 10 2023;3(1):141. doi:10.1038/s43856-023-00370-1
22. Yang D, Wei J, Xiao D, et al. PediatricsGPT: Large Language Models as Chinese Medical Assistants for Pediatric Applications. *ArXiv*. 2024;abs/2405.19266
23. Estill J. The application of large language models in pediatrics and medical research—Revolution or risk? *Pediatric Discovery*. 2023;1(3):e39. doi:https://doi.org/10.1002/pdi3.39
24. Alhur A. Redefining Healthcare With Artificial Intelligence (AI): The Contributions of ChatGPT, Gemini, and Co-pilot. *Cureus*. Apr 2024;16(4):e57795. doi:10.7759/cureus.57795
25. Rossettini G, Bargeri S, Cook C, et al. Accuracy of ChatGPT-3.5, ChatGPT-4o, Copilot, Gemini, Claude, and Perplexity in advising on lumbosacral radicular pain against clinical practice guidelines: cross-sectional study. *Front Digit Health*. 2025;7:1574287. doi:10.3389/fdgth.2025.1574287
26. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. *Health Care Sci*. Aug 2023;2(4):255-263. doi:10.1002/hcs2.61
27. Hatem R, Simmons B, Thornton JE. A Call to Address AI "Hallucinations" and How Healthcare Professionals Can Mitigate Their Risks. *Cureus*. Sep 2023;15(9):e44720. doi:10.7759/cureus.44720
28. Jones-Jordan LA, Walline JJ, Mutti DO, et al. Gas permeable and soft contact lens wear in children. *Optom Vis Sci*. Jun 2010;87(6):414-20. doi:10.1097/OPX.0b013e3181dc9a04
29. Wang C, Gallo RE, Fleisher L, Miller SM. Literacy assessment of family health history tools for public health prevention. *Public Health Genomics*. 2011;14(4-5):222-37. doi:10.1159/000273689
30. Kochanek K, Skarzynski H, Jedrzejczak WW. Accuracy and Repeatability of ChatGPT Based on a Set of Multiple-Choice Questions on Objective Tests of Hearing. *Cureus*. May 2024;16(5):e59857. doi:10.7759/cureus.59857
31. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res*. Jun 30 2023;25:e47479. doi:10.2196/47479
32. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Front Artif Intell*. 2023;6:1237704. doi:10.3389/frai.2023.1237704