DOI: 10.4274/jcrpe.galenos.2026.2026-1-15

Original Article

A Comparative Assessment of Large Language Models in Congenital Hypothyroidism: Reliability, Quality and Readability

Barsal Çetiner E and Singin B. Quality of LLMs in Congenital Hypothyroidism

Ebru Barsal Çetiner, Berna Singin
University of Health Sciences Türkiye, Antalya City Hospital, Department of Pediatric Endocrinology, Antalya, Türkiye

What is already known on this topic?
Large language model (LLM)–based chatbots are increasingly used by patients to obtain medical information. Previous studies have shown variable reliability, quality, and readability of LLM-generated health content, and most materials exceed the recommended sixth-grade reading level for patient education.

What this study adds?
This study is the first to evaluate four LLMs in the context of congenital hypothyroidism using parent-centered questions, assessing reliability, quality, readability, and source accuracy. Although all models demonstrated good levels of reliability and quality, ChatGPT-5.2 showed superior overall performance compared with the others. These findings suggest that, as LLMs continue to evolve, they hold increasing potential to generate more reliable and readable health information.

Abstract
Objective: To comparatively evaluate the reliability, quality, and readability of responses generated by widely used large language model (LLM)–based chatbots to congenital hypothyroidism (CH)–related patient questions.
Methods: Forty CH frequently asked questions (FAQs), derived from clinician-reviewed patient education resources, were submitted under standardized conditions (December 2025) to ChatGPT-4, ChatGPT-5.2, Gemini, and Copilot. The modified DISCERN (mDISCERN) instrument was used to assess reliability, whereas the Global Quality Score (GQS) was used to evaluate quality. Readability was evaluated using Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Coleman–Liau Index (CLI), and Simple Measure of Gobbledygook (SMOG). Scores were compared using Friedman tests with Bonferroni-corrected post hoc analyses.
Results: Median mDISCERN scores were 5.0 for ChatGPT-4, ChatGPT-5.2, and Gemini, and 4.0 for Copilot. Median GQS scores were 5.0 for ChatGPT-4, ChatGPT-5.2, and Gemini, and 4.0 for Copilot. Differences among models were significant for both mDISCERN and GQS ($p<0.001$), with ChatGPT-5.2 outperforming others in key pairwise comparisons. Readability differed significantly across all indices (all $p<0.001$). ChatGPT-5.2 demonstrated the highest FRE and lowest FKGL, whereas Gemini produced the most complex text. However, all models exceeded the recommended sixth-grade reading level.
Conclusion: LLM-based chatbots generated generally moderate-to-high quality CH information, but readability remains suboptimal for patient education. ChatGPT-5.2 showed the best overall performance. LLM outputs may support patient information needs but should complement, not replace, clinician-provided counseling.
Keywords: Artificial intelligence, ChatGPT, congenital hypothyroidism, Copilot, Google Gemini, Large language models

Ebru Barsal Çetiner, MD, University of Health Sciences Türkiye, Antalya City Hospital, Department of Pediatric Endocrinology, Antalya, Türkiye
ebrubarsalcetiner@gmail.com
https://orcid.org/0000-0002-1888-919X

29.01.2026
10.04.2026

Epub: 21.04.2026

Introduction

Primary congenital hypothyroidism (CH) is the most common congenital endocrine disorder, with an estimated incidence of approximately 1 in 1,000–3,000 live births worldwide. If left untreated, CH may lead to severe and irreversible intellectual disability; however, this adverse outcome can largely be prevented through neonatal screening programs and early initiation of treatment. In countries where newborn screening programs are effectively implemented, the majority of patients with CH demonstrate neurodevelopmental outcomes within normal limits (1,2,3,4,5,6,7).
Abnormal findings on newborn screening necessitate confirmatory biochemical testing to establish or exclude the diagnosis of hypothyroidism. Measurement of thyrotropin (thyroid-stimulating hormone, TSH) together with free thyroxine (free T4), or alternatively total T4 and triiodothyronine (T3) uptake, is recommended for this purpose. The presence of elevated serum TSH levels accompanied by low free T4 concentrations confirms the diagnosis of primary hypothyroidism and requires urgent initiation of treatment (8). Oral levothyroxine is the treatment of choice, and both the timing and dosage of thyroid hormone replacement are critical determinants of clinical outcomes (9,10). In term infants, the recommended initial dose is 10–15 µg/kg/day, a range that has been associated with optimal neurocognitive outcomes, normal growth, and improved school performance (10,11).
Nowadays, patients—and particularly parents of newborns—have long been actively using the internet to access health-related information. Previous reports indicate that approximately 90% of adults use the internet, and nearly 75% search for health-related information before seeking medical care, highlighting the importance of evaluating the accuracy and readability of online medical content (12). Medical content is disseminated to broad audiences through digital and social media platforms such as Google, Facebook, and Twitter (13). In recent years, the use of artificial intelligence (AI) technologies in the field of healthcare has increased rapidly (14). AI refers to the ability of computer

systems to perform functions that typically require human intelligence, including decision-making, learning from experience, natural language understanding, and problem-solving.

In addition to traditional citation metrics, alternative metrics such as Altmetric scores provide valuable insight into the dissemination and societal impact of scientific publications. Bibliometric analyses not only illustrate the historical development of a research field but also identify highly interactive studies that shape academic visibility. Recent bibliometric studies highlight the growing role of digital engagement in the dissemination of medical knowledge, emphasizing the importance of evaluating online information sources in contemporary healthcare environments (13).

In this context, AI–driven large language models (LLMs**)** have emerged as novel and easily accessible sources of information for individuals seeking health-related knowledge. AI-based chatbots are capable of interacting with patients, answering questions, and providing basic medical information (15). Chat Generative Pre-Trained Transformer (ChatGPT) version 3.5 was released in 2022, rapidly gained a large user base, and was subsequently followed by more advanced versions (16). In addition to ChatGPT, other LLM-based chatbots, such as Microsoft Copilot and Google Gemini, have also been developed. These models were selected because they represent the most widely used and publicly accessible LLM-based chatbots at the time of the study and have been frequently evaluated in previous healthcare-related research, allowing comparability with existing literatüre.

Artificial intelligence enables patients to access health-related information easily; however, health literacy plays a crucial role in patient understanding and engagement, and the readability, reliability, and quality of this information are of critical importance (17).The National Institutes of Health (NIH), the American Medical Association (AMA), and the United States Department of Health and Human Services recommend that web-based patient education materials be written at or below a sixth-grade reading level (18,19,20,21). In addition, LLMs may occasionally cite non-existent sources or generate biased or inaccurate information, which raises concerns regarding patient safety, particularly in conditions that may lead to irreversible neurodevelopmental outcomes (22). Improved patient knowledge regarding disease mechanisms and treatment has been shown to enhance adherence to medical recommendations and improve clinical outcomes (23).

The aim of this study was to conduct a comparative evaluation of the responses generated by four AI-based chatbots—ChatGPT-4, ChatGPT-5.2, Gemini, and Copilot—to frequently asked questions (FAQs) related to congenital hypothyroidism, with respect to readability, reliability, and quality.

1. Materials and Methods

1.1. Study Design

This study was designed as a cross-sectional analytical study evaluating the reliability, quality, and readability of responses generated by AI–based LLMs regarding congenital hypothyroidism.

1.2. Question Sources and Initial Screening

Questions related to CH were developed using patient education content from internationally recognized, reliable, evidence-based websites that are reviewed by clinicians, including the Cleveland Clinic, Mayo Clinic, and the United Kingdom National Health Service (NHS). These sources were selected because they are widely regarded as trustworthy in patient and caregiver education and include questions that are frequently asked by patients and their families.

Initially, 60 questions related to congenital hypothyroidism were identified. Questions that were repetitive, highly similar in wording, overlapping in meaning, or not directly related to CH were excluded through a screening process. Following this refinement, 40 questions were selected for the final analysis. The complete list of questions is provided in the Supporting Information section.

1.3. Question Categorization

The final set of questions was categorized into six clinically meaningful domains reflecting the topics most frequently sought by parents of children with CH. These domains included basic information; symptoms and clinical features; diagnosis and screening; treatment and monitoring; risks, side effects, and complications; and recovery and outlook.

1.4. Artificial Intelligence Models and Interaction Procedure

The selected questions were submitted to multiple LLM-based chatbot platforms, including ChatGPT-4 and ChatGPT-5.2 (free and paid versions; OpenAI; December 2025 and December 2025), Gemini (free version; Google; November 2025), and Copilot (Microsoft; December 2025), all of which were publicly accessible at the time of the study. All evaluations were conducted in December 2025 using identical prompts and standardized conditions. All searches were performed using a web browser in incognito mode without logging into any personal accounts to minimize personalization bias.

To ensure that each response was generated independently and to prevent contextual memory bias, the conversation history was cleared prior to each question, and a new chat session was initiated. To assess response consistency, the same set of questions was resubmitted to each chatbot one week later under the same conditions. No additional prompts, follow-up questions, response regeneration commands, or clarifications were used, except for requesting references when they were not initially provided by the chatbot.

All responses and cited references were recorded and stored for subsequent analysis. The existence, accessibility, and academic credibility of the cited sources were systematically verified and documented. All cited references were systematically verified using PubMed, Google Scholar, CrossRef, and official journal websites. A reference was classified as fabricated if it could not be identified in these databases or if inconsistencies were detected in authorship, journal name, publication year, volume, page numbers, or DOI information. In addition, references that were retrievable but unrelated to the topic or that did not support the statements made in the chatbot response were classified as inaccurate citations. All references were independently reviewed by two pediatric endocrinologists, and disagreements were resolved by consensus. Source usage and citation behavior were incorporated into the modified DISCERN (mDISCERN) and Global Quality Scale (GQS) assessments, and misleading, fabricated, or non-academic references were systematically identified and recorded.

1.5. Expert Evaluation Process

All chatbot responses were independently evaluated by two pediatric endocrinologists with clinical experience in the management of CH. In cases of disagreement regarding scoring, the responses were re-assessed by a third pediatric endocrinologist, and a final decision was reached by consensus. Inter-rater agreement exceeded 0.80 (Cohen's κ), indicating excellent agreement.

1.6. Reliability Assessment

The mDISCERN instrument was used to assess reliability. This scale consists of five criteria, with each criterion scored as 1 if fulfilled and 0 if not fulfilled. Higher total scores (out of five) indicate greater reliability. The reliability and validity of the DISCERN instruments have been previously established (24,25). The mDISCERN scale evaluates the following five criteria using a yes/no format: clear statement of aims; reliability of information sources; balance and absence of bias; provision of additional sources of information; and discussion of uncertainties.

1.7. Quality Assessment

The quality of the chatbot responses was evaluated using the GQS, which has been applied in similar studies (26,27). GQS is a five-point Likert scale designed to assess the usability, quality, and flow of online health information. A score of 1 represents the lowest quality, whereas a score of 5 indicates the highest quality. Scores of 2 reflect low quality with limited usefulness, 3 indicate moderate quality with limited usefulness, and 4 represent good quality and usefulness (Table 1).

1.8. Readability Assessment

The readability of the responses generated by the chatbots was analyzed using multiple established readability indices to evaluate textual complexity and the required reading level. These indices included the Flesch Reading Ease (FRE), Flesch–Kincaid Grade Level (FKGL),

Gunning Fog Index (GFI), Coleman–Liau Index (CLI), and the Simple Measure of Gobbledygook (SMOG). Readability scores were calculated using an online tool with automated computation functions (28).

The FRE score ranges from 0 to 100, with lower scores indicating more difficult text. Scores between 0 and 30 correspond to very difficult texts requiring college-level reading skills; scores between 31 and 50 indicate difficult texts appropriate for grades 13–16; scores between 51 and 60 represent relatively difficult texts at the 10th–12th grade level; scores between 61 and 70 indicate plain English suitable for grades 8–9; scores between 71 and 80 correspond to fairly easy texts at the 7th-grade level; scores between 81 and 90 indicate easy texts appropriate for the 6th-grade level; and scores between 91 and 100 represent very easy texts that can be understood by an average 11-year-old student.

The FKGL score represents the grade level required to understand a text, with scores of 10 or higher indicating that the material is appropriate for readers at the high school level or above. According to recommendations from the AMA and the NIH, patient education materials should be written at a sixth-grade reading level or lower (18,19,20,21).

The CLI measures the reading level corresponding to grade levels in the United States. The SMOG score indicates the number of years of education required to understand a text. In the GFI, which evaluates textual complexity based on sentence length and the proportion of long words, scores above 12 indicate more difficult texts.

Acceptable readability thresholds were defined as an FRE score of ≥80 and ≤6 for the other four indices. Materials exceeding these thresholds were considered more difficult to read than the levels recommended for the general population (23,29,30,31,32).

1.9. Statistical Analysis

Statistical analyses were performed using SPSS software (IBM Corp., Armonk, NY). Continuous and ordinal variables were assessed for normality using the Shapiro–Wilk test. As the outcome scores did not follow a normal distribution, results were summarized as median and interquartile range (IQR). Differences in scores across the compared chatbot models were evaluated using the Friedman test. When the Friedman test indicated a statistically significant difference, post hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test with Bonferroni correction. A Bonferroni-corrected p value of <0.008 was considered statistically significant. Effect sizes for the Friedman tests were calculated using Kendall's W and interpreted as small (≈0.1), moderate (≈0.3), and large (≥0.5). This approach was used to complement p-values and to provide information on the magnitude of differences among models.

1.10.Ethics

This study did not involve human participants or patient-level data. All evaluated responses were obtained from publicly accessible artificial intelligence platforms. Therefore, ethics committee approval was not required.

2. Results

The response performance of ChatGPT-4, ChatGPT-5.2, Gemini, and Copilot was evaluated using CH-related FAQs grouped into six domains. These domains comprised Basic Information; Symptoms and Clinical Features; Diagnosis and Screening; Treatment and Monitoring; Risks, Side Effects, and Complications; and Recovery and Outlook.

The reliability of the LLMs was assessed using mDISCERN. The median mDISCERN score was 5.0 (4.0–5.0) for ChatGPT-4, 5.0 (5.0–5.0) for ChatGPT-5.2, 5.0 (4.0–5.0) for Gemini, and 4.0 (3.0–4.0) for Copilot (Table 2).

The quality of the responses was evaluated using GQS. The median GQS score was 5.0 (4.0–5.0) for both ChatGPT-4 and Gemini, 5.0 (5.0–5.0) for ChatGPT-5.2, and 4.0 (3.0–4.0) for Copilot (Table 2).

The readability of the responses to the FAQs was evaluated using multiple indices. The highest FRE score was observed for ChatGPT-5.2 at 57.2 (39.4–66.8), whereas the lowest FRE score was recorded for Gemini at 38.2 (31.1–46.8). The lowest FKGL score was also obtained for ChatGPT-5.2 at 8.4 (7.0–12.0), while the highest FKGL score was found for ChatGPT-4 at 13.3 (11.9–14.5). ChatGPT-5.2 demonstrated the lowest SMOG, GFI, and CLI scores, whereas Gemini had the highest values for these indices (Table 3).

Gemini and Copilot provided references and direct links to the cited sources after each response. ChatGPT-4 and ChatGPT-5.2 did not provide sources by default but supplied references when explicitly requested; ChatGPT-5.2 included hyperlinks, whereas ChatGPT-4 did not. Regarding the accuracy of the cited sources, ChatGPT-5.2 achieved a rate of 100%, ChatGPT-4 and Gemini each demonstrated an accuracy of 95%, and Copilot showed an accuracy rate of 60%.

All LLMs provided additional information beyond the direct answers and indicated what further details they could offer upon request. In addition, ChatGPT-5.2 presented brief summary sections for parents (e.g., "short answer for parents") for some questions. The responses generated by ChatGPT-4 were generally longer than those of the other models.

2.1.Reliability and Quality

All LLMs differed significantly in terms of mDISCERN scores ($p < 0.001$). The Friedman test yielded $\chi^2(3)=22.653$ ($p<0.001$), with a Kendall's W of 0.19, indicating a small-to-moderate effect size. In pairwise comparisons, significant differences were observed between ChatGPT-5.2 and ChatGPT-4 and between ChatGPT-5.2 and Copilot ($p = 0.002$ and $p < 0.001$, respectively) (Table 2). ChatGPT-5.2 achieved higher reliability scores than the other models.

With respect to content quality, GQS scores also differed significantly among all LLMs ($p < 0.001$). The effect size was small-to-moderate ($\chi^2(3)=22.393$, $p<0.001$; Kendall's W=0.19). In pairwise comparisons, the GQS score of ChatGPT-5.2 was significantly higher than those of the other three models ($p = 0.001$, $p = 0.001$, and $p < 0.001$, respectively) (Table 2). No significant differences in reliability or quality scores were observed across the different question categories.

2.2.Readability

Significant differences were observed among the AI models across all readability indices (SMOG, FKGL, GFI, CLI, and FRE; all $p < 0.001$). Effect size analysis demonstrated large effects for FKGL (W=0.59), SMOG (W=0.55), CLI (W=0.58), and FRE (W=0.49), and a very large effect for GFI (W=0.86), indicating substantial differences in textual complexity across models(Table 3). ChatGPT-5.2 demonstrated significantly higher FRE scores and significantly lower FKGL, SMOG, GFI, and CLI scores compared with all other models (Table 4), indicating superior readability.

In pairwise comparisons among ChatGPT-4, Gemini, and Copilot, mixed results were observed depending on the index. No significant difference was found between ChatGPT-4 and Copilot for the CLI score, or between Gemini and Copilot for the SMOG score ($p = 0.624$) (Table 4).

FRE: Significant differences were observed among the models ($p < 0.001$). Copilot was more readable than ChatGPT-4 and Gemini ($p = 0.002$ and $p < 0.001$, respectively). ChatGPT-5.2 was more readable than all other models (all $p < 0.001$).

FKGL: Significant differences were found among the models ($p < 0.001$). Copilot was more readable than ChatGPT-4 and Gemini, and Gemini was more readable than ChatGPT-4 (all $p < 0.001$). ChatGPT-5.2 demonstrated significantly better readability than all other models (all $p < 0.001$).

SMOG: Significant differences were detected across all models ($p < 0.001$). ChatGPT-4 had lower SMOG scores than Gemini and Copilot ($p < 0.001$ and $p = 0.002$, respectively). ChatGPT-5.2 showed lower SMOG scores than all other models (all $p < 0.001$).

GFI: GFI scores ranked from lowest to highest as ChatGPT-5.2, ChatGPT-4, Copilot, and Gemini, with all pairwise comparisons being statistically significant.

CLI: Based on CLI scores, ChatGPT-4 and Copilot were more readable than Gemini (both $p < 0.001$). ChatGPT-5.2 had significantly lower CLI scores than all other models ($p = 0.001$ vs. ChatGPT-4; $p < 0.001$ vs. Gemini and Copilot) (Tables 3 and 4).

3. Discussion

In this study, the quality, reliability, and readability of responses provided by ChatGPT-4, ChatGPT-5.2, Gemini, and Copilot to the most frequently asked questions related to CH were evaluated. To the best of our knowledge, this is the first study to compare the responses of four different LLMs to CH-related FAQs.

As is well known, CH is diagnosed during the neonatal period, most commonly through heel-prick blood screening. If early treatment is not initiated, it can lead to irreversible intellectual disability and developmental delay, making it a major source of anxiety for parents of newborns (8). For this reason, many parents seek information on this condition through LLMs. These systems have been reported to assist healthcare professionals in areas such as disease diagnosis, treatment planning, prognostic assessment, and public health management, and they may also influence patient decision-making in healthcare (33). By comparing the quality, reliability, and readability of LLMs, the present study provides insight into their suitability for use by parents.

In our study, the reliability and quality of all LLMs were found to be in the moderate-to-high range, with statistically significant differences observed among them. Ranked from lowest to highest, the models were Copilot, ChatGPT-4 and Gemini, and ChatGPT-5.2. ChatGPT-5.2 was significantly more reliable than ChatGPT-4 and Copilot and demonstrated higher quality than ChatGPT-4, Gemini, and Copilot. Consistent with our findings, a previous study evaluating ChatGPT, Perplexity, ChatSonic, and Microsoft Bing AI reported that the information quality of the responses was moderate to high (34). Gül et al. (14) found lower mDISCERN scores for ChatGPT and Gemini and higher scores for Perplexity. Another study reported that Gemini achieved higher GQS scores compared with other chatbots (35). The superior performance of ChatGPT-5.2 in our study may be attributed to its concise and accurate responses and the high accuracy of the sources it provided, while Gemini's provision of references and direct links alongside its answers likely contributed to its relatively high scores.

Recent studies have shown that the accuracy, quality, and clinical appropriateness of LLM responses depend largely on the clarity and specificity of user prompts. The study by Sarangi and Mondal (36) also emphasized that LLMs such as ChatGPT, Google Bard, and Microsoft Bing perform better when prompted with clear and well-defined queries. Accordingly, we formulated our questions to be concise and unambiguous, drawing on internationally recognized, evidence-based, and clinician-reviewed patient education resources. Within the scope of the validated assessment tools used in this study, all evaluated LLMs demonstrated generally high levels of reliability and quality.We also evaluated the readability level of the responses in our study. In the United States, the average literacy level corresponds to approximately the 7th–8th grade; however, according to the AMA, health education materials should be written at the 6th-grade level. This recommendation is based on the fact that patients' comprehension decreases when they are dealing with illness and psychological stress, and therefore even complex medical conditions should be explained in very simple language (27). Nevertheless, previous studies have shown that a substantial proportion of online patient education materials exceed the recommended readability levels, which is considered inappropriate from a public health perspective (19,20,21,24).

While statistically significant differences were observed across models, effect size analysis showed that reliability and quality differences were small-to-moderate, whereas readability differences were large to very large. This suggests that although overall content quality was relatively comparable among models, substantial variability exists in textual complexity, which may have meaningful implications for patient comprehension and health literacy.

In our study, the responses generated by ChatGPT-5.2 were found to correspond approximately to a 9th–10th grade reading level, whereas those of Copilot corresponded to a 12th–14th grade level, ChatGPT-4 to a 13th–14th grade level, and Gemini to a 14th–16th grade level. ChatGPT-5.2 was statistically significantly more readable than all other LLMs. Pairwise comparisons among ChatGPT-4, Gemini, and Copilot yielded variable significant differences depending on the readability index used.

Momenaei et al. (37) reported that understanding ChatGPT's responses on retinal disease surgery required a university-level education. Similarly, another study found that responses provided by ChatGPT, Bard, and Microsoft Bing Chat to palliative care–related questions were written at approximately a 10th-grade reading level (38). Although studies directly comparing ChatGPT, Gemini, Copilot, and particularly ChatGPT-5.2 in terms of readability are limited, existing evidence, consistent with our findings, indicates that the readability of LLM-generated content generally exceeds the recommended 6th-grade level. This raises important concerns about how such content can be made more accessible for individuals with lower educational attainment.

Previous studies have demonstrated that ChatGPT versions can reduce readability levels when provided with specific instructions (39,40,41). These findings suggest that incorporating tailored prompts aimed at simplifying language could enhance readability in future applications. The superior readability of ChatGPT-5.2 observed in our study may also be attributed to its more advanced language model architecture compared with the other LLMs.

In a study evaluating the knowledge levels of caregivers of children with CH, insufficient knowledge was identified as a major barrier to effective follow-up. It was emphasized that healthcare professionals providing information about this condition—one of the leading preventable causes of intellectual and developmental disability—should use clear, simple, and understandable language (42). Education is a key component of disease management (43), and studies have shown that providing patients and caregivers with personalized information improves adherence to medical recommendations and leads to better health outcomes (44).

In this context, the use of LLMs by caregivers, in addition to the education provided by healthcare professionals, has become increasingly common with the recent expansion of AI-based applications. Although the use of LLMs is known to enhance access to healthcare information, concerns remain regarding the potential for misleading content, variability in quality, and readability levels that may exceed those appropriate for the general population (14,45). Therefore, AI-based tools should be used cautiously, and consultation with healthcare professionals should be encouraged when necessary. Consistent with this, all LLMs evaluated in our study included warning statements advising users to consult a physician or noting that the information provided should not be used as a substitute for medical decision-making. Consistent with previous research, digital platforms such as YouTube and web-based resources represent major sources of health information for patients; however, studies have demonstrated considerable variability in the readability, reliability, and quality of such content, emphasizing the importance of ongoing evaluation of online health information (18,46)

4. Limitations

In this study, the analysis was limited to English-language responses, as English is the most commonly used language in general online information seeking. Therefore, the findings cannot be directly generalized to content generated in other languages. In addition, the use of a single readability calculator may have introduced minor variability in readability estimates, although the tool employed has been widely used in previous studies (35)Furthermore, the findings are based on chatbot responses obtained in December 2025; given the continuous updating of LLMs, these results may change over time. LLM-generated responses are not fully deterministic and may vary across sessions or over time due to model updates and probabilistic generation mechanisms. Therefore, exact reproducibility of responses cannot be fully guaranteed.

5. Strengths

This study represents the first comprehensive evaluation of multiple LLMs specifically in congenital hypothyroidism. The use of validated and widely accepted assessment tools, standardized prompts, and expert evaluation enhances the methodological robustness and objectivity of the findings, enabling a reliable comparison across models.

6. Conclusion

This study showed that although all four chatbots produced CH-related content with moderate to good reliability and quality, ChatGPT-5.2 outperformed the others in reliability, quality, and readability, despite overall readability exceeding the recommended sixth-grade level. CH

is a condition that can be detected early through neonatal heel-prick screening and remains a major public health concern. Therefore, the potential of AI-based tools to provide accurate, understandable, and reliable information to patients and caregivers is of great importance. To minimize the risk of misinformation, LLMs should expand their knowledge bases, rely on credible academic sources, and generate content with readability levels that better match the general population's health literacy. Nevertheless, regardless of how advanced LLMs become, they cannot replace face-to-face medical consultations and clinical evaluation between patients and physicians.

Acknowledgements
We would like to thank the pediatric endocrinologists who scored the performance of the large language models in answering frequently asked patient questions about CH
Authors' contributions
Concept: E.B.Ç.; Design: E.B.Ç.; B.S.; Data Collection: E.B.Ç.; B.S.; Analysis: E.B.Ç.; B.S.; Writing: E.B.Ç.; B.S.; Critical Review: E.B.Ç.; B.S.; Approval: E.B.Ç.; B.S.
Funding
The authors did not receive any funding for the research.
Conflict of Interest: The authors declare no conflict of interest.
Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request. The complete list of FAQs included in the analysis is available in the Supporting Information.
Ethics Statement
This study did not involve human participants or patient-level data. All evaluated responses were obtained from publicly accessible artificial intelligence platforms. Therefore, ethics committee approval was not required.

References
1. Kurinczuk JJ, Bower C, Lewis B, Byrne G. Congenital hypothyroidism in Western Australia 1981–1998. J Paediatr Child Health. 2002;38:187-191. doi:10.1046/j.1440-1754.2002.00812.x
2. Hinton CF, Harris KB, Borgfeld L, Drummond-Borg M, Eaton R, Lorey F, Therrell BL. Trends in incidence rates of congenital hypothyroidism related to select demographic factors. Pediatrics. 2010;125(Suppl 2):S37-S47. doi:10.1542/peds.2009-1975D
3. Waller DK, Anderson JL, Lorey F, Cunningham GC. Risk factors for congenital hypothyroidism: an investigation of infant's birth weight, ethnicity, and gender. Teratology. 2000;62:36-41. doi:10.1002/1096-9926(200007)62:1<36::AID-TERA8>3.0.CO;2-W
4. Tuli G, Munarin J, Tessaris D, Matarazzo P, Einaudi S, de Sanctis L. Incidence of primary congenital hypothyroidism and relationship between diagnostic categories and associated malformations. Endocrine. 2021;71:122-129. doi:10.1007/s12020-020-02370-w
5. Deladoëy J, Ruel J, Giguère Y, Van Vliet G. Is the incidence of congenital hypothyroidism really increasing? A 20-year retrospective population-based study in Québec. J Clin Endocrinol Metab. 2011;96:2422-2429. doi:10.1210/jc.2011-1073
6. Danner E, Niuro L, Huopio H, Niinikoski H, Viikari L, Kero J, et al. Incidence of primary congenital hypothyroidism over 24 years in Finland. Pediatr Res. 2022;93:649-653. doi:10.1038/s41390-022-02118-4
7. McGrath N, Hawkes C, McDonnell C, Cody D, O'Connell SM. Incidence of congenital hypothyroidism over 37 years in Ireland. Pediatrics. 2018;142:e20181199. doi:10.1542/peds.2018-1199
8. van Trotsenburg P, Stoupa A, Léger J, Rohrer T, Peters C, Fugazzola L, Cassio A. Congenital hypothyroidism: a 2020–2021 consensus guidelines update. Thyroid. 2021;31:387-419. doi:10.1089/thy.2020.0333
9. Selva KA, Harper A, Downs A, Blasco PA, Lafranchi SH. Neurodevelopmental outcomes in congenital hypothyroidism: comparison of initial T4 dose and time to reach target T4 and TSH. J Pediatr. 2005;147:775-780. doi:10.1016/j.jpeds.2005.07.024
10. Rose SR, Wassner AJ, et al. Congenital hypothyroidism: screening and management. Pediatrics. 2023;151:e2022060420. doi:10.1542/peds.2022-060420
11. Aleksander PE, Brückner-Spieler M, Stoehr AM, Lankes E, Kühnen P, Schnabel D, Ernert A, Stäblein W, Craig ME, Blankenstein O, Grüters A, Krude H. Mean high-dose l-thyroxine treatment is efficient and safe to achieve a normal IQ in young adult patients with congenital hypothyroidism. *J Clin Endocrinol Metab.* 2018;103:1459-1469. doi:10.1210/jc.2017-01937
12. Gunduz ME, Matis GK, Ozduran E, Hancı V. Evaluating the readability, quality, and reliability of online patient education materials on spinal cord stimulation. *Turk Neurosurg* 34.4 (2024): 588-599
13. Bağcıer F, Yurdakul O, Özduran E. Top 100 cited articles on ankylosing spondylitis. *Biruni Univ Open Access Repository.* 2020. Available from: https://openaccess.biruni.edu.tr/xmlui/handle/20.500.12445/1858
14. Gül Ş, Erdemir İ, Hancı V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: Assessment of readability, reliability, and quality of ChatGPT and BARD. *Med (Baltimore).* 2024;103. 18: e38009. doi:10.1097/MD.0000000000038009
15. Hopkins JM, Logan J, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. *JNCI Cancer Spectr.* 2023;7:pkad010. doi:10.1093/jncics/pkad010
16. Zhang S, Liau ZQG, Tan KLM, Chua WL. Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. *Knee Surg Relat Res.* 2024;36:15. doi:10.1186/s43019-024-00218-5
17. Ozduran E, Hancı V, Erkin Y. Evaluating the readability, quality and reliability of online patient education materials on chronic low back pain. *National Medical Journal of India* 37.3 (2024).
18. Erkin Y, Hancı V, Özduran E. Evaluation of the reliability and quality of YouTube videos as a source of information for transcutaneous electrical nerve stimulation. *PeerJ.* 2023;11:e15412. doi:10.7717/peerj.15412
19. Özduran E, Hancı V. Evaluating the readability, quality, and reliability of online information on Sjögren's syndrome. *Indian J Rheumatol.* 2023;18:16-25. doi:10.4103/injr.injr_56_22
20. Özduran E, Hancı V. Evaluating the readability, quality and reliability of online information on Behçet's disease. *Reumatismo.* 2022;74.2: 49-60. doi:10.4081/reumatismo.2022.1495
21. Özduran E, Büyükçoban S. Evaluating the readability, quality and reliability of online patient education materials on post-COVID pain. *PeerJ.* 2022;10:e13686. doi:10.7717/peerj.13686
22. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15(2):e35179. doi:10.7759/cureus.138667
23. Özbek İ, Hancı V, of EÖTJ, 2025 undefined. Digital Guidance: Quality and Readability Analysis of Artificial Intelligence-Generated Spondyloarthropathy Texts. *Turkish Journal of Osteoporosis* (2025). 31(1):12-18
24. Erkin Y, Hancı V, Özduran E. Evaluating the readability, quality and reliability of online patient education materials on transcutaneous electrical nerve stimulation (TENS). *Med (Baltimore).* 2023;102.16:e33529. doi: 10.1097/MD.0000000000033529
25. Ladhar S, Koshman SL, Yang F, Turgeon R. Evaluation of online written medication educational resources for people living with heart failure. *CJC Open.* 2022;4.10:858-865. doi:10.1016/j.cjco.2022.07.004

26. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect.* 2004;7.2:165-175. doi:10.1111/j.1369-7625.2004.00273.x
27. Avra TD, Le M, Hernandez S, Thure K, Ulloa JG. Readability assessment of online peripheral artery disease education materials. *J Vasc Surg.* 2022;76.6:1728-1732 doi:10.1016/j.jvs.2022.07.022
28. Simpson D. The Readability Test Tool. Available from: https://www.readabilityformulas.com
29. Pohl N, Derector E, Rivlin M, Patel A, Graham B. A quality and readability comparison of artificial intelligence and popular health website education materials for common hand surgery procedures. *Ann Burns Fire Disasters.* 2024;43.3:101723 doi:10.1016/j.hansur.2024.101723
30. Lee B, Dixon E, Wales DP. Evaluation of reading level of result letters sent to patients from an academic primary care practice. *Health Serv Res Manag Epidemiol.* 2023;10:23333928231172142. doi:10.1177/23333928231172142
31. Onder E, Ensari E. ChatGPT-4o's performance on pediatric vesicoureteral reflux. *J Pediatr Urol.* 2025.21(2),504-509 doi: 10.1016/j.jpurol.2024.12.002
32. Kara M, Ozduran E, Kara MM, Özbek İC, Hancı V. Evaluating the readability, quality, and reliability of responses generated by ChatGPT, Gemini, and Perplexity on the most commonly asked questions about Ankylosing spondylitis." *Plos one* 20.6 (2025): e0326351.
33. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health Serv Res Manag Epidemiol.* 2024;11:23333928241234863. doi:10.1177/23333928241234863
34. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur Urol.* 2024;85.1:13-16. doi: 10.1016/j.eururo.2023.07.004
35. Özduran E, Akkoc İ, Büyükçoban S, Erkin Y, Hancı V. Readability, reliability and quality of responses generated by ChatGPT, Gemini, and Perplexity for the most frequently asked questions about pain. *Med (Baltimore).* 2025;104.11;e41780. doi:10.1097/MD.0000000000041780
36. Sarangi P, Mondal H. Response generated by large language models depends on the structure of the prompt. *Indian J Radiol Imaging.* 2024;34:574-575. doi:10.1055/s-0044-1782165
37. Momenaei, B., Wakabayashi, T., Shahlaee, A., Durrani, A. F., Pandit, S. A., Wang, K., ... & Kuriyan, A. E. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retina.* 2023;7:862-868. doi:10.1016/j.oret.2023.05.022
38. Kim MJ, Admane S, Chang YK, Shih KK, Reddy A, Tang M, La Cruz M, Taylor TP. Chatbot performance in defining and differentiating palliative care, supportive care, hospice care. *J Pain Symptom Manage.* 2024;67:e381-e391. doi: 10.1016/j.jpainsymman.2024.01.008
39. Foster BJ, Mitsnefes M, Dahhou X, Zhang X, Laskin BL. Changes in excess mortality from end-stage renal disease in the United States from 1995 to 2013. *Clin J Am Soc Nephrol.* 2018;13:91-99. doi: 10.2215/CJN.04330417
40. Garcia Valencia OA, Thongprayoon C, Miao J, Suppadungsuk S, Krisanapan P, Craici IM, Cheungpasitporn W. Empowering inclusivity: improving readability of living kidney donation information with ChatGPT. *Front Digit Health.* 2024;6:1366967. doi:10.3389/fdgth.2024.1366967
41. Zaki HA, Mai M, Abdel-Megid H, Liew SQR, Kidanemariam S, Omar AS, Tiwari U, Hamze J. Using ChatGPT to improve readability of interventional radiology procedure descriptions. *Cardiovasc Intervent Radiol.* 2024;47:1134-1141. doi:10.1007/s00270-024-03803-z
42. Brito LNS, de Andrade CLO, de Aragão Dantas Alves C. Adhesion to treatment by children with congenital hypothyroidism: knowledge of caregivers in Bahia State, Brazil. *Rev Paul Pediatr.* 2021;39:e2020074. doi:10.1590/1984-0462/2021/39/2020074
43. Wolf MS, Gazmararian JA, Baker DW. Health literacy and functional health status among older adults. *Arch Intern Med.* 2005;165:1946-1952. doi:10.1001/archinte.165.17.1946
44. Grippaudo FR, Nigrelli S, Patrignani A, Ribuffo D, Tognini S, La Mastra C. Quality of the information provided by ChatGPT for patients in breast plastic surgery: are we already in the future? *JPRAS Open.* 2024;40:99-105. doi: 10.1016/j.jpra.2024.02.001
45. Association of Women's Health, Obstetric and Neonatal Nurses. Health literacy. Available from: http://lib.ncfh.org/pdfs/6617.pdf
46. Özduran, Erkan. ""Bel Ağrısı" ile İlgili Türkçe İnternet Kaynaklı Hasta Eğitim Materyallerinin Okunabilirliklerinin Değerlendirilmesi." *Dokuz Eylül Üniversitesi Tıp Fakültesi Dergisi* 36.2 (2022): 135-150.

Table 1. Global quality scale (GQS) criteria.

1. Poor quality, poor flow of the site, most information missing, not at all useful for patients
2. Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients
3. Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients
4. Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for Patients
5. Excellent quality and excellent flow, very useful for patients

Table 2. Comparison of mDISCERN and GQS scores across AI models

AI Model	mDISCERN Median (Q1–Q3)	vs ChatGPT-5 (p)	vs Gemini (p)	GQS Median (Q1–Q3)	vs ChatGPT-5 (p)	vs Gemini (p)
ChatGPT-4	5.0 (4.0-5.0)	**0.002**	0.655	5.0 (4.0-5.0)	**0.001**	0.721
ChatGPT-5.2	5.0 (5.0–5.0)	Reference	-	5.0 (5.0-5.0)	Reference	-
Gemini	5.0 (4.0–5.0)	0.008	Reference	5.0 (4.0-5.0)	**0.001**	Reference
Copilot	4.0 (3.0–4.0)	**<0.001**	0.036	4.0 (3.0-4.0)	**<0.001**	0.021

Values are presented as median (interquartile range [Q1–Q3]).Overall differences among AI models were assessed using the Friedman test. Post-hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction ($p < 0.008$). Asterisks (*) indicate statistically significant differences.

Table 3. Comparison of readability scores across AI models

Readability index	ChatGPT-4	Gemini	Copilot	ChatGPT-5.2	*p*
FRE	41.2 (36.3–48.9)	38.2 (31.1–46.8)	46.9 (39.8–52.8)	57.2 (39.4–66.8)	**<0.001**
FKGL	13.3 (11.9–14.5)	11.9 (11.0–13.9)	10.2 (8.8–12.8)	8.4 (7.0–12.0)	**<0.001**

SMOG	12.6 (11.3–13.9)	13.9 (12.4–15.8)	13.2 (12.0–14.6)	9.8 (7.4–13.5)	**<0.001**
GFI	14.0 (12.8–15.8)	15.8 (13.6–17.2)	14.6 (13.4–16.3)	10.2 (7.9–14.9)	**<0.001**
CLI	12.9 (11.8–14.7)	14.2 (12.4–16.8)	12.9 (11.8–14.1)	9.1 (7.3–13.9)	**<0.001**

Values are presented as median (interquartile range [Q1–Q3]). Comparisons across AI models were performed separately for each readability index using the Friedman test. Statistically significant differences ($p < 0.05$).
FRE: Flesch Reading Ease; FKGL: Flesch–Kincaid Grade Level; SMOG: Simple Measure of Gobbledygook; GFI: Gunning Fog Index; CLI: Coleman–Liau Index.

Table 4. Post-hoc Wilcoxon signed-rank test results for readability indices

Comparison	**FRE *p***	**FKGL *p***	**SMOG *p***	**GFI *p***	**CLI *p***
ChatGPT-4 vs Gemini	0.013	**<0.001**	**<0.001**	**<0.001**	**<0.001**
ChatGPT-4 vs Copilot	**0.002**	**<0.001**	**0.002**	**<0.001**	0.624
Gemini vs Copilot	**<0.001**	**<0.001**	0.014	**0.001**	**<0.001**
ChatGPT-4 vs ChatGPT-5.2	**<0.001**	**<0.001**	**<0.001**	**0.002**	**0.001**
Gemini vs ChatGPT-5.2	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Copilot vs ChatGPT-5.2	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

All p-values were obtained using the Wilcoxon signed-rank test with Bonferroni correction. Adjusted significance level was set at $p < 0.008$.
FRE: Flesch Reading Ease; FKGL: Flesch–Kincaid Grade Level; SMOG: Simple Measure ofGobbledygook; GFI: Gunning Fog Index; CLI: Coleman–Liau Index.