

Article

Reliability and Quality of AI-Generated Information on Newborn Screening Tests: A Comparative Analysis of ChatGPT and Gemini

Ece Nilüfer ^{1,*} , Meserret Aslan ¹  and Şehma Şen ²¹ Department of Midwifery, Faculty of Health Sciences, Istanbul Atlas University, Istanbul 34403, Türkiye; meserret.aslan@atlas.edu.tr² Department of Nursing English, Faculty of Health Sciences, Istanbul Atlas University, Istanbul 34403, Türkiye; sehma.sen@atlas.edu.tr

* Correspondence: ece.nilufer@atlas.edu.tr

Abstract

Background/Objectives: The increasing use of artificial intelligence (AI) chatbots for obtaining health-related information has raised concerns regarding the quality and reliability of the information they provide. This study aimed to compare the quality and reliability of responses generated by ChatGPT free tier (GPT-4o, with GPT-4.1 mini as the fallback model after the usage limit) and Gemini 2.5 Flash (Google, free version) regarding newborn screening tests. **Methods:** A total of 31 questions were developed based on international and national newborn screening guidelines and were posed to both chatbots. Responses were independently evaluated by two researchers using the DISCERN instrument and the Global Quality Score (GQS), and inter-rater reliability was assessed. Descriptive statistics and non-parametric tests were used to compare chatbot performance. **Results:** Both evaluators assigned significantly higher DISCERN total scores to Gemini than to ChatGPT free tier. For Evaluator 1, the mean DISCERN scores were 48.8 ± 10.2 for Gemini and 42.4 ± 5.9 for ChatGPT free tier ($p < 0.001$); for Evaluator 2, the corresponding scores were 53.7 ± 8.9 and 42.4 ± 5.8 , respectively ($p < 0.001$). For the GQS ratings, Evaluator 1 rated Gemini significantly higher than ChatGPT free tier (3.6 ± 0.6 vs. 3.0 ± 0.5 , $p < 0.001$), whereas Evaluator 2 found no statistically significant difference between the two chatbots (3.3 ± 1.3 vs. 3.4 ± 0.8 , $p = 0.877$). Inter-rater reliability for DISCERN scores was excellent for ChatGPT and moderate for Gemini, whereas agreement for GQS ratings was low for both chatbots. **Conclusions:** Gemini demonstrated consistently higher DISCERN scores than ChatGPT free tier; however, its superiority in overall GQS ratings was not consistently supported across the two evaluators. Neither chatbot consistently achieved the highest levels of information quality. AI chatbots should therefore be considered supplementary sources of health information rather than substitutes for healthcare professionals or official health information resources.



Academic Editor: Francisco José García-Peñalvo

Received: 2 June 2026

Revised: 30 June 2026

Accepted: 9 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.**Keywords:** newborn screening; artificial intelligence; reliability; ChatGPT; consumer health information

1. Introduction

Neonatal screening refers to the totality of diagnostic procedures administered in the early stages of the postnatal period which, when performed with optimum timing and methodology, play a critical role in the early diagnosis and prevention of pathologies

carrying risks of mortality and severe morbidity [1]. Neonatal screening, a fundamental public health activity conducted shortly after birth, prioritizes the early detection of serious yet treatable conditions. Recognized by the WHO as a highly effective intervention [2], this practice stands as a historically proven and successful program. Many countries possess institutionalized screening systems, and an estimated 38 to 40 million infants are screened annually for at least one disabling disorder. The early detection and treatment provided enable affected infants to achieve their maximum health and developmental potential [3]. Traditional Neonatal Screening, initiated in the 1960s with the aim of detecting serious and treatable diseases, is grounded in the principles established by Wilson and Junger [4]. Although the number of diseases included within the scope of screening has increased over the years, the standards governing the selection process have remained valid [5].

Neonatal metabolic screening was first initiated in the early 1960s by Robert Guthrie through the collection of blood samples on filter paper (Guthrie card) for Phenylketonuria screening via the bacterial inhibition assay. This program was subsequently expanded in 1974 with the inclusion of screening for Congenital Hypothyroidism [6]. Although screening protocols exhibit global heterogeneity, the 'National Newborn Screening Program' in Turkey was mandated in 2006 regarding Phenylketonuria and Congenital Hypothyroidism. The panel was subsequently expanded to include Biotinidase Deficiency in 2008 and Cystic Fibrosis in 2015. Furthermore, pilot studies for Congenital Adrenal Hyperplasia screening were initiated in selected provinces in 2017. In addition to metabolic assays, the comprehensive screening regimen encompasses audiological and ophthalmological evaluations, as well as ultrasonographic assessment for Developmental Dysplasia of the Hip (DDH) [7].

Artificial Intelligence (AI), which mimics human intelligence, and its subset Machine Learning (ML), are among the most important technologies of today [8]. AI enables machines to perform complex tasks on their own. Machine Learning, on the other hand, allows computers to learn and improve by themselves using data, without being coded step by step [9,10]. AI is reshaping the way we process information; by leveraging capabilities such as learning, reasoning, and problem-solving, it strengthens decision-making frameworks and pioneers' transformation across numerous domains of life [11].

In the rapidly developing landscape of AI assistants, comparative analyses of ChatGPT and Gemini (formerly Google Bard) have become a critical focal point, especially regarding the quality and reliability of their responses to medical and health-related questions [9,10]. Recent studies in the literature assess the accuracy and effectiveness of the AI models in question in answering inquiries within different medical branches [12–15]. The integration of AI in healthcare encounters multifaceted obstacles, such as data privacy, algorithmic precision, ethical compliance, societal acceptance, and clinical adaptation [16,17]. These constraints complicate the development of highly accurate and reliable systems. Sensitivities regarding data privacy limit access to information, while potential biases in algorithms can lead to erroneous clinical outcomes [16,18].

With the increasing public availability of generative AI, individuals have begun to use conversational AI tools as sources of health information. In a cross-sectional survey of 2406 respondents in the United States, 21.5% reported using ChatGPT to obtain online health information, commonly to determine whether professional consultation was needed or to explore alternative treatment options. The conversational format, ease of access, and ability to provide immediate explanations may therefore make AI chatbots appealing to individuals seeking understandable information about health-related concerns [19].

Newborn screening represents a context in which parents may have urgent and complex information needs. A screen-positive result does not necessarily establish a definitive diagnosis and may require repeat or confirmatory testing, leaving families un-

certain while awaiting follow-up. Parents may also encounter unfamiliar names of rare disorders and require understandable information about false-positive results, prognosis, treatment options, follow-up procedures, and available support resources. Previous qualitative research has shown that parents may experience distress following the communication of a positive screening result, encounter difficulties in obtaining reliable and timely information, and turn to internet searches, particularly when there is a delay between notification of the result and consultation with a knowledgeable healthcare professional. However, online information concerning rare disorders may be confusing, outdated, or frightening [20]. In addition, parental newborn screening information materials do not always provide sufficient explanations regarding false-positive and false-negative results or subsequent follow-up processes [21]. Under these circumstances, freely accessible AI chatbots may become an appealing supplementary source of information for families seeking rapid clarification.

Although several studies have evaluated the performance of AI chatbots in different medical specialties, evidence regarding the quality and reliability of AI-generated information on newborn screening tests remains limited. Newborn screening represents a unique area of preventive healthcare that requires clear, reliable, and understandable communication for families. Therefore, evaluating the quality of information provided by widely used AI chatbots is of particular importance.

The aim of this study was to compare the quality and reliability of responses generated by ChatGPT free tier and Gemini regarding newborn screening tests using the DISCERN instrument and the Global Quality Score (GQS). The findings may contribute to a better understanding of the potential role of AI chatbots as supplementary sources of health information in the field of newborn screening.

2. Materials and Methods

2.1. AI-Based Chatbots Included in the Study

This study evaluated responses generated by two widely used artificial intelligence (AI) chatbots: ChatGPT free tier, using GPT-4o with GPT-4.1 mini as the fallback model after the free usage limit (OpenAI, San Francisco, CA, USA), and Gemini 2.5 Flash free version (Google LLC, Mountain View, CA, USA). The freely accessible versions were selected because the study aimed to evaluate AI-generated health information available to members of the general public without requiring a paid subscription. This approach was intended to reflect real-world access conditions for individuals seeking information about newborn screening tests, rather than to compare premium or frontier AI models.

ChatGPT is a large language model designed to generate human-like text responses based on user prompts, while Gemini is a multimodal AI system capable of processing and generating information across multiple formats. Both platforms are increasingly used as sources of medical and health-related information, making their evaluation particularly relevant in the context of newborn screening education.

Data collection was conducted on 18 July 2025. All 31 questions were submitted to both chatbots on the same date using the versions publicly available at that time. ChatGPT was accessed through its free tier, and Gemini 2.5 Flash was accessed through the free version of the Gemini web interface. To minimize contextual bias, each question was entered into a new chat session with no prior conversation history retained. Because each session was independent, all questions were phrased to explicitly refer to newborn screening tests, ensuring that sufficient context was provided without relying on previous prompts. All questions were submitted in Turkish, and the chatbot responses were evaluated in their original Turkish form. The questions and selected response excerpts presented in the manuscript were translated into English for publication.

2.2. Development of Questions Related to Newborn Screening Tests

A total of 31 questions related to newborn screening tests were developed by two researchers (E.N. and M.A.) based on recommendations and information provided in guidelines published by the World Health Organization (WHO) and the Turkish Ministry of Health. The questions were designed to cover key aspects of newborn screening, including the purpose of screening programs, screening procedures, diseases included in screening panels, interpretation of results, follow-up processes, ethical considerations, and the roles of healthcare professionals.

To ensure consistency in data collection, the same set of questions was submitted to both chatbots without modification. Each question was entered independently, and responses were recorded for subsequent evaluation. The complete list of questions is presented in Table 1, and selected examples of the chatbot responses are presented in Table 2. The DISCERN and GQS scoring criteria, the independent scores assigned by the two evaluators, and the inter-rater reliability results are provided in Supplementary Tables S1–S3.

Table 1. Questions Related to Newborn Screening Tests.

Questions Related to Newborn Screening Tests	
1.	What are newborn screening tests?
2.	What are the primary purposes of newborn screening tests?
3.	Which newborns are eligible for newborn screening tests?
4.	In which healthcare settings are newborn screening tests performed?
5.	On which day are newborn screening tests conducted?
6.	Is parental consent required for newborn screening tests?
7.	What are the common errors observed in newborn screening tests?
8.	Which diseases are included in newborn screening programs?
9.	What is phenylketonuria screening, and how is it performed?
10.	What is congenital hypothyroidism screening, and how is it performed?
11.	What is biotinidase deficiency screening, and how is it performed?
12.	What is cystic fibrosis screening, and how is it performed?
13.	What is congenital adrenal hyperplasia screening, and how is it performed?
14.	What is spinal muscular atrophy (SMA) screening, and how is it performed?
15.	What is newborn eye screening, and how is it performed?
16.	How is the red reflex test performed?
17.	What is newborn hearing screening, and how is it performed?
18.	What is developmental dysplasia of the hip (DDH) screening, and how is it performed?
19.	How is the heel-prick blood sampling procedure performed?
20.	How long does it take to receive newborn screening test results?
21.	What steps are taken following a positive newborn screening test result?
22.	Do newborn screening tests provide a definitive diagnosis?

Table 1. *Cont.*

Questions Related to Newborn Screening Tests	
23.	Can newborn screening tests produce false-positive results?
24.	What is the prognosis of conditions detected early by newborn screening tests?
25.	For which conditions in newborn screening tests is the timing of treatment initiation critical?
26.	What complications may develop if the condition detected in newborn screening tests is left untreated?
27.	How are newborn screening test results communicated to parents?
28.	Is the treatment process of conditions detected by newborn screening tests covered by the government?
29.	How can families exercise their right to refuse newborn screening tests?
30.	How is the confidentiality of newborn screening test results protected?
31.	What are the roles and responsibilities of healthcare professionals in the newborn screening test process?

Table 2. Some examples of the chatbots' responses to the questions posed.

	What Is Congenital Hypothyroidism Screening, and How Is It Performed?	What Complications May Develop If the Condition Detected in Newborn Screening Tests Is Left Untreated?
ChatGPT	Congenital hypothyroidism is a condition where the thyroid gland doesn't work properly from birth. If left untreated, this condition can lead to serious problems like intellectual disability and growth delays. The test is done using a heel prick blood sample from the baby, and it measures the levels of thyroid hormones.	Untreated diseases can result in developmental delays, neurological problems, physical disabilities, and even death in infants.
Gemini	Congenital hypothyroidism is a condition in newborns where the thyroid gland does not produce enough thyroid hormone. If not diagnosed and treated early, mental and physical developmental delays may occur. How it is done: It is done by measuring the thyroid-stimulating hormone (TSH) level in a blood sample taken from the heel.	If left untreated, various serious complications may develop depending on the screened diseases: • Mental retardation • Physical developmental delay • Neurological damage • Organ damage • Serious infections • Life-threatening metabolic crises • Muscle weakness and paralysis • Permanent disability or death

2.3. Evaluation Criteria

The responses generated by ChatGPT free tier and Gemini were evaluated using two widely recognized instruments for assessing the quality of health information: the DISCERN instrument and the Global Quality Score (GQS).

DISCERN is a standardized tool developed to assess the reliability and quality of written consumer health information regarding treatment choices. The instrument consists of 16 items, each rated on a 5-point Likert scale ranging from 1 (low quality) to 5 (high quality). Higher total scores indicate greater reliability and overall quality of the information

provided. The DISCERN instrument has been widely used in studies evaluating online health information and AI-generated medical content [22].

The Global Quality Score (GQS), originally introduced by Bernard et al. [23], is a five-point rating system used to assess the overall quality, flow, completeness, ease of use, and usefulness of online health information. Scores range from 1, indicating poor-quality information with substantial information gaps and little usefulness, to 5, indicating excellent-quality, comprehensive, and highly useful information. Higher scores reflect better overall informational quality and greater value for users.

All 62 chatbot responses were independently evaluated by two researchers using the DISCERN instrument and the Global Quality Score (GQS). The complete DISCERN evaluation criteria and the five-point GQS scoring descriptions are presented in Supplementary Table S1. The evaluators completed the assessments separately and were blinded to each other's ratings. The independent DISCERN total scores and GQS ratings assigned by each evaluator are presented in Supplementary Table S2. Inter-rater reliability was assessed before the comparative analyses. No consensus scores were generated. Comparative analyses between ChatGPT and Gemini were conducted separately for each evaluator to preserve the independence of the ratings.

2.4. Statistical Analysis

Descriptive statistics, including the mean, standard deviation, median, minimum and maximum values, and interquartile ranges (25th–75th percentiles), were calculated separately for each evaluator and chatbot for the DISCERN total scores and GQS ratings. The normality of the paired score differences between the two chatbots was assessed using the Shapiro–Wilk test. Because the assumption of normality was not met, the DISCERN and GQS ratings of ChatGPT and Gemini were compared separately for each evaluator using the Wilcoxon signed-rank test. Effect sizes were calculated as $r = |Z|/N$, where N represents the number of non-zero paired differences. Inter-rater reliability for the total DISCERN scores was assessed using a two-way random-effects, absolute-agreement, single-measures intraclass correlation coefficient (ICC (2,1)) with a 95% confidence interval. Agreement between the two evaluators for the ordinal GQS ratings was assessed using quadratic-weighted Cohen's kappa with a 95% confidence interval. No consensus or averaged scores were generated, and all comparative analyses were conducted separately for each evaluator. A p -value of less than 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS Statistics version 29.0.

2.5. Ethical Statement

Ethical approval was not required for this study because it did not involve human participants, animals, patient data, or identifiable personal information. The study exclusively evaluated responses generated by publicly available artificial intelligence chatbots. No personal, clinical, or sensitive data were collected, processed, or analyzed during the study.

3. Results

Inter-rater reliability results for the independent DISCERN and GQS ratings are presented in Supplementary Table S3. For ChatGPT responses, inter-rater reliability for the DISCERN total scores was excellent (ICC (2,1) = 0.972, 95% CI: 0.943–0.986, $p < 0.001$), whereas it was moderate for Gemini responses (ICC (2,1) = 0.553, 95% CI: 0.199–0.768, $p < 0.001$). Agreement between the two evaluators for the GQS ratings was low and not statistically significant for both ChatGPT (quadratic-weighted $\kappa = 0.182$, 95% CI: −0.127 to 0.490, $p = 0.199$) and Gemini (quadratic-weighted $\kappa = 0.125$, 95% CI: −0.109 to

0.359, $p = 0.328$). Comparisons of the DISCERN and GQS ratings between ChatGPT and Gemini according to the two independent evaluators are presented in Table 3.

Table 3. Comparison of DISCERN and GQS Ratings Between ChatGPT and Gemini According to Independent Evaluators.

Measure	Evaluator	ChatGPT Mean $\pm$ SD	Median (Q1–Q3)	Min–Max	Gemini Mean $\pm$ SD	Median (Q1–Q3)	Min–Max	Z	p-Value	Effect Size r
DISCERN total score	Evaluator 1	42.4 $\pm$ 5.9	41 (37–48)	33–57	48.8 $\pm$ 10.2	47 (39–59)	34–65	−3.748	<0.001	0.71
DISCERN total score	Evaluator 2	42.4 $\pm$ 5.8	42 (38–47)	33–57	53.7 $\pm$ 8.9	55 (45–61)	37–66	−4.311	<0.001	0.79
GQS	Evaluator 1	3 $\pm$ 0.5	3 (3–3)	2–4	3.6 $\pm$ 0.6	4 (3–4)	2–4	−3.662	<0.001	0.86
GQS	Evaluator 2	3.4 $\pm$ 0.8	3 (3–4)	2–5	3.3 $\pm$ 1.3	4 (3–4)	1–5	−0.155	0.877	0.03

Note: Comparisons were performed separately for each evaluator using the Wilcoxon signed-rank test because the same 31 questions were submitted to both chatbots. Effect size was calculated as $r = |Z| / \sqrt{N}$, where N represents the number of non-zero paired differences. GQS, Global Quality Score; Q1, 25th percentile; Q3, 75th percentile; SD, standard deviation.

Comparisons of the DISCERN and GQS ratings between ChatGPT and Gemini according to the two independent evaluators are presented in Table 3. Evaluator 1 assigned significantly higher DISCERN total scores to Gemini than to ChatGPT free tier (48.8 $\pm$ 10.2 vs. 42.4 $\pm$ 5.9; $Z = -3.748$, $p < 0.001$), with a large effect size ($r = 0.71$). Similarly, Evaluator 2 assigned significantly higher DISCERN total scores to Gemini than to ChatGPT free tier (53.7 $\pm$ 8.9 vs. 42.4 $\pm$ 5.8; $Z = -4.311$, $p < 0.001$), also with a large effect size ($r = 0.79$). For the GQS ratings, Evaluator 1 rated Gemini significantly higher than ChatGPT free tier (3.6 $\pm$ 0.6 vs. 3.0 $\pm$ 0.5; $Z = -3.662$, $p < 0.001$), with a large effect size ($r = 0.86$). In contrast, Evaluator 2 found no statistically significant difference between the GQS ratings of Gemini and ChatGPT free tier (3.3 $\pm$ 1.3 vs. 3.4 $\pm$ 0.8; $Z = -0.155$, $p = 0.877$), and the effect size was negligible ($r = 0.03$). Thus, although both evaluators consistently assigned higher DISCERN scores to Gemini, the comparative GQS findings differed between the evaluators.

4. Discussion

This study compared the quality and reliability of information generated by ChatGPT free tier and Gemini regarding newborn screening tests using the DISCERN instrument and the Global Quality Score (GQS). Both evaluators assigned significantly higher DISCERN total scores to Gemini than to ChatGPT free tier, indicating a consistent advantage for Gemini in the dimensions assessed by the DISCERN instrument. In contrast, the comparative GQS findings were evaluator-dependent: Evaluator 1 rated Gemini significantly higher than ChatGPT free tier, whereas Evaluator 2 found no statistically significant difference between the two chatbots. Therefore, the present findings support a consistent advantage for Gemini in DISCERN scores but do not demonstrate uniform superiority across both assessment tools. These findings suggest that differences in the design and capabilities of large language models may influence the quality of health information provided to users. The results are consistent with a growing body of literature reporting improved performance of newer AI models in delivering more comprehensive and user-oriented medical information [15]. For example, Ozduran et al. (2025) found that Gemini outperformed ChatGPT in terms of GQS ratings when responding to commonly asked questions regarding pain, indicating superior informational quality despite ongoing readability issues [24]. The idea that model evolution may be associated with improvements in content quality is further supported by Grilo et al. (2025), who found that more sophisticated GPT models produced higher-quality responses than GPT-3.5 in radiotherapy-related patient

queries [25]. The present results extend these observations to the preventive public health setting of newborn screening, which is characterized by probabilistic outcomes, ethical sensitivity, and the need for clear communication with families.

Because AI chatbot platforms are updated frequently, the present findings should be interpreted as a time-specific snapshot rather than as a permanent comparison between the two systems. Changes in the underlying models, training data, system instructions, safety mechanisms, and response-generation processes may alter the outputs produced in response to identical questions. Therefore, exact replication of this study later may yield different responses and scores. The observed differences between ChatGPT and Gemini should consequently be interpreted as reflecting the versions available on 18 July 2025 and should not be generalized to current or future versions of either platform. The inter-rater reliability findings also warrant consideration. Agreement between the two evaluators for the DISCERN total scores was excellent for ChatGPT free tier and moderate for Gemini. In contrast, agreement for the GQS ratings was low and not statistically significant for either chatbot. This limited agreement may reflect the greater subjectivity of the GQS as a single-item global assessment and is consistent with the evaluator-dependent comparative GQS findings. Accordingly, the GQS results should be interpreted cautiously, whereas the DISCERN findings provide more consistent evidence of differences between the two chatbots.

Gemini achieved higher DISCERN scores than ChatGPT free tier across most of the evaluated questions, suggesting that its responses were generally perceived as higher in quality and reliability. Similar findings have been reported in previous studies evaluating AI-generated health information in different medical specialties [26,27]. Nevertheless, neither chatbot consistently achieved scores corresponding to the highest quality levels. This finding indicates that, although AI chatbots may provide useful health information, there remains room for improvement in the overall quality and reliability of their responses.

A notable strength of this study is its focus on newborn screening tests, an area that has received limited attention in previous evaluations of AI-generated health information. Unlike many medical topics that focus primarily on diagnosis or treatment, newborn screening involves preventive healthcare, early detection strategies, follow-up procedures, and communication of results to families. Therefore, providing clear and reliable information in this context is particularly important. The findings of the present study contribute to the growing literature on AI-generated health information by examining chatbot performance within a preventive healthcare framework that requires accurate and understandable communication [28].

Despite the superior performance of Gemini, several challenges associated with AI-generated health information remain evident. One of the most reported limitations in the literature is the lack of explicit citation of reliable sources, which may reduce users' ability to verify the information provided [25,26,29]. In addition, previous studies have highlighted concerns regarding the readability and accessibility of AI-generated medical content, particularly for individuals with limited health literacy [24,27]. These concerns are especially relevant in the context of newborn screening, where parents must understand complex health information to make informed decisions and comply with follow-up recommendations.

From an ethical and clinical perspective, AI chatbots should be considered supplementary sources of health information rather than substitutes for healthcare professionals or official public health resources. Although Gemini consistently received higher DISCERN scores and was rated more favorably on the GQS by one evaluator, AI-generated responses may still contain omissions, inaccuracies, or oversimplifications that could lead to misunderstandings if used without professional guidance. Therefore, information provided

by AI chatbots should be interpreted cautiously and verified using trusted healthcare sources [15,17].

This study highlights the importance of evaluating AI-generated health information within specific healthcare contexts. Although AI chatbot performance has been investigated across various medical specialties, newborn screening presents unique informational and ethical considerations that require particular attention. Future research should explore the performance of newer AI models, assess the impact of AI-generated information on parental understanding and decision-making, and examine the integration of evidence-based references into chatbot responses. Such investigations may provide further insight into the potential role of AI chatbots in supporting health education and communication.

In summary, Gemini received significantly higher DISCERN total scores than ChatGPT free tier from both evaluators, whereas its advantage in GQS ratings was supported by only one evaluator. Neither chatbot consistently achieved the highest levels of information quality, highlighting the need for continued improvement in AI-generated health information. Although AI chatbots may serve as useful supplementary sources of information for families seeking information about newborn screening, they should not replace healthcare professionals or official health information resources. Future developments should focus on improving information quality, transparency, readability, and evidence-based response generation to support the safe and effective use of AI technologies in health communication.

Limitations

This study has several limitations. First, only two AI chatbots and their freely accessible versions were evaluated; therefore, the findings may not be generalizable to newer, premium, or other commercially available AI models. Second, the evaluation was based on the DISCERN and GQS instruments, which assess information quality and reliability but do not directly measure the factual accuracy of chatbot responses. Third, inter-rater reliability for the DISCERN total scores was moderate for Gemini, while agreement for the GQS ratings was low and not statistically significant for either chatbot. These findings indicate that some evaluations, particularly the single-item GQS ratings, may have been influenced by evaluator subjectivity and should therefore be interpreted cautiously. Fourth, all chatbot responses were generated at a single time point. Because commercial AI chatbots are frequently updated, changes in the underlying models, system instructions, safety mechanisms, and response-generation processes may affect the outputs produced for identical prompts. Therefore, exact replication of the present findings may be difficult, and the results should be interpreted as reflecting the versions available on 18 July 2025 rather than the current or future performance of either platform. Finally, the study focused exclusively on general questions regarding newborn screening tests and may not reflect chatbot performance in other areas of healthcare or in response to disease-specific parental concerns. Future studies should evaluate the quality, reliability, and factual accuracy of responses to questions that parents are likely to ask following a positive newborn screening result, particularly regarding confirmatory testing, treatment, prognosis, follow-up procedures, and available support resources.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/informatics13070115/s1>. Supplementary Table S1: DISCERN Instrument and Global Quality Score Scoring Criteria; Supplementary Table S2: Independent DISCERN Total Scores and GQS Ratings Assigned by the Two Evaluators; Supplementary Table S3: Inter-Rater Reliability of Independent DISCERN and GQS Ratings.

Author Contributions: Conceptualization, E.N. and M.A.; methodology, E.N., M.A. and Ş.Ş.; validation, E.N., M.A. and Ş.Ş.; formal analysis, E.N.; investigation, E.N., M.A. and Ş.Ş.; data curation, E.N. and Ş.Ş.; writing—original draft preparation, E.N.; writing—review and editing, M.A. and Ş.Ş.;

supervision, M.A.; project administration, E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it did not involve human participants, animals, patient data, or identifiable personal information. The study exclusively evaluated responses generated by publicly available artificial intelligence chatbots.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available within the article and its Supplementary Materials.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Therrell, B.L.; Padilla, C.D.; Loeber, J.G.; Kneisser, I.; Saadallah, A.; Borrajo, G.J.; Adams, J. Current status of newborn screening worldwide: 2015. *Semin. Perinatol.* **2015**, *39*, 171–187. [CrossRef] [PubMed]
2. Tam, P.-Y.I.; Padilla, C.D.; Zlotkin, S.; Ayede, A.I.; Banu, T.; Kayita, J.; Khanna, R.; Rao, S.P.; Siddeeg, K.; Walani, S.; et al. The 77th World Health Assembly resolution calling for newborn screening, diagnosis, and management of birth defects: Moving towards action in low-income and middle-income countries. *Lancet Glob. Health* **2024**, *12*, e1754–e1755. [CrossRef] [PubMed]
3. Webster, D.; Gaviglio, A.; Khan, A.H.; Baker, M.; Cheillan, D.; Chabraoui, L.; Abdoh, G.; Cabello, J.; Giugliani, R.; Platis, D.; et al. ISNS general guidelines for neonatal bloodspot screening 2025. *Int. J. Neonatal Screen.* **2025**, *11*, 45. [CrossRef] [PubMed]
4. Wilson, J.M.G.; Jungner, G. *Principles and Practice of Screening for Disease*; World Health Organization: Geneva, Switzerland, 1968. Available online: <https://apps.who.int/iris/handle/10665/37650> (accessed on 4 February 2026).
5. Downie, L.; Halliday, J.; Lewis, S.; Amor, D.J. Principles of genomic newborn screening programs: A systematic review. *JAMA Netw. Open* **2021**, *4*, e2114336. [CrossRef] [PubMed]
6. Guthrie, R.; Susi, A. A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics* **1963**, *32*, 338–343. [CrossRef]
7. Altunhan, H.; Yılmaz, F.H. Yenidoğanın değerlendirilmesi ve yenidoğan taramaları. *Turk. Klin. J. Fam. Med.-Spec. Top.* **2018**, *9*, 28–32.
8. Fattah, F.H.; Salih, A.M.; Salih, A.M.; Asaad, S.K.; Ghafour, A.K.; Bapir, R.; Abdalla, B.A.; Othman, S.; Ahmed, S.M.; Hasan, S.J.; et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: A scoping review. *Front. Digit. Health* **2025**, *7*, 1482712. [CrossRef] [PubMed]
9. Abbas, Y.N.; Mahmood, Y.M.; Hassan, H.A.; Hamad, D.Q.; Hasan, S.J.; Omer, D.A.; Kakamad, S.H.; HamaSalih, H.M.; Hassan, M.N.; Rahim, H.M.; et al. Role of ChatGPT and Google Bard in the diagnosis of psychiatric disorders: A cross-sectional study. *Barw Med. J.* **2023**, *1*, 14–19. [CrossRef]
10. Hiwa, D.S.; Abdalla, S.S.; Muhialdeen, A.S.; Hamasalih, H.M.; Karim, S.O. Assessment of nursing skill and knowledge of ChatGPT, Gemini, Microsoft Copilot, and Llama: A comparative study. *Barw Med. J.* **2024**, *2*, 3–6. [CrossRef]
11. Xu, L.; Sanders, L.; Li, K.; Chow, J.C.L. Chatbot for health care and oncology applications using artificial intelligence and machine learning: Systematic review. *JMIR Cancer* **2021**, *7*, e27850. [CrossRef] [PubMed]
12. Salih, A.M.; Mohammed, N.A.; Mahmood, Y.M.; Hassan, S.J.; Namiq, H.S.; Ghafour, A.K.; Asaad, S.K.; Kakamad, S.H.; Abdullah, F.; Hussein, K.F.; et al. ChatGPT insight and opinion regarding the controversies in neurogenic thoracic outlet syndrome: A case-based study. *Barw Med. J.* **2023**, *1*, 2–4. [CrossRef]
13. Fuchs, A.; Trachsel, T.; Weiger, R.; Eggmann, F. ChatGPT's performance in dentistry and allergy immunology assessments: A comparative study. *SWISS Dent. J. SSO-Sci. Clin. Top.* **2024**, *134*, 1–17. [CrossRef] [PubMed]
14. Masalkhi, M.; Ong, J.; Waisberg, E.; Lee, A.G. Google DeepMind's Gemini AI versus ChatGPT: A comparative analysis in ophthalmology. *Eye* **2024**, *38*, 1412–1417. [CrossRef] [PubMed]
15. Wei, Q.; Yao, Z.; Cui, Y.; Wei, B.; Jin, Z.; Xu, X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *J. Biomed. Inform.* **2024**, *151*, 104620. [CrossRef] [PubMed]
16. Sharma, D.; Kaushal, S.; Kumar, H.; Gainder, S. Chatbots in healthcare: Challenges, technologies and applications. In Proceedings of the 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST), Delhi, India, 9–10 December 2022; pp. 1–6. [CrossRef]
17. Ray, P.P.; Majumder, P. The potential of ChatGPT to transform healthcare and address ethical challenges in artificial intelligence-driven medicine. *J. Clin. Neurol.* **2023**, *19*, 509–511. [CrossRef] [PubMed]

18. Farhud, D.D.; Zokaei, S. Ethical issues of artificial intelligence in medicine and healthcare. *Iran. J. Public Health* **2021**, *50*, i–v. [[CrossRef](#)] [[PubMed](#)]
19. Ayo-Ajibola, O.; Davis, R.J.; Lin, M.E.; Riddell, J.; Kravitz, R.L. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: Cross-sectional questionnaire study. *J. Med. Internet Res.* **2024**, *26*, e55138. [[CrossRef](#)] [[PubMed](#)]
20. Conway, M.; Vuong, T.T.; Hart, K.; Rohrwasser, A.; Eilbeck, K. Pain points in parents' interactions with newborn screening systems: A qualitative study. *BMC Pediatr.* **2022**, *22*, 167. [[CrossRef](#)] [[PubMed](#)]
21. Ijzebrink, A.; van Dijk, T.; Franková, V.; Loeber, G.; Kožich, V.; Henneman, L.; Jansen, M. Informing parents about newborn screening: A European comparison study. *Int. J. Neonatal Screen.* **2021**, *7*, 13. [[CrossRef](#)] [[PubMed](#)]
22. 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* **1999**, *53*, 105–111. [[CrossRef](#)] [[PubMed](#)]
23. Bernard, A.; Langille, M.; Hughes, S.; Rose, C.; Leddin, D.; Veldhuyzen van Zanten, S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **2007**, *102*, 2070–2077. [[CrossRef](#)] [[PubMed](#)]
24. Ozduran, E.; Akkoc, I.; 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. *Medicine* **2025**, *104*, e41780. [[CrossRef](#)] [[PubMed](#)]
25. Grilo, A.; Marques, C.; Corte-Real, M.; Carolino, E.; Caetano, M. Assessing the quality and reliability of ChatGPT's responses to radiotherapy-related patient queries: Comparative study with GPT-3.5 and GPT-4. *JMIR Cancer* **2025**, *11*, e63677. [[CrossRef](#)] [[PubMed](#)]
26. Ghanem, Y.K.; Rouhi, A.D.; Al-Houssan, A.; Saleh, Z.; Moccia, M.C.; Joshi, H.; Dumon, K.R.; Hong, Y.; Spitz, F.; Joshi, A.R.; et al. Dr. Google to Dr. ChatGPT: Assessing the content and quality of artificial intelligence-generated medical information on appendicitis. *Surg. Endosc.* **2024**, *38*, 2887–2893. [[CrossRef](#)] [[PubMed](#)]
27. Zhou, M.; Pan, Y.; Zhang, Y.; Song, X.; Zhou, Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. *Int. J. Med. Inform.* **2025**, *198*, 105871. [[CrossRef](#)] [[PubMed](#)]
28. Piazza, D.; Martorana, F.; Curaba, A.; Sambataro, D.; Valerio, M.R.; Firenze, A.; Pecorino, B.; Scollo, P.; Chiantera, V.; Scibilia, G.; et al. The consistency and quality of ChatGPT responses compared to clinical guidelines for ovarian cancer: A Delphi approach. *Curr. Oncol.* **2024**, *31*, 2796–2804. [[CrossRef](#)] [[PubMed](#)]
29. Cherrez-Ojeda, I.; Faytong-Haro, M.; Alvarez-Muñoz, P.; Larco, J.I.; de Arruda Chaves, E.; Rojo, I.; Moncayo, C.V.; Ramon, G.D.; Rodas-Valero, G.; Kocatürk, E.; et al. How accurate are ChatGPT-4 responses in chronic urticaria? A critical analysis with information quality metrics. *World Allergy Organ. J.* **2025**, *18*, 101071. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.