Original Article

DOI: 10.4274/haseki.galenos.2026.63383
Med Bull Haseki 2026;64(4):229-236

Expert Evaluation of Large Language Models for Caregiver Counseling in Pediatric Seizures: A Comparative Study of ChatGPT, Gemini, and DeepSeek

Sema Bozkaya Yilmaz[1], Gunce Basarir[2]

[1]*University of Health Sciences Türkiye, Bursa City Hospital, Clinic of Pediatric Neurology, Bursa, Türkiye*

[2]*University of Health Sciences Türkiye, Istanbul Haseki Training and Research Hospital, Clinic of Pediatric Neurology, Istanbul, Türkiye*

Abstract

Aim: Caregiver-facing large language model (LLM) responses to pediatric seizure questions must provide accurate, complete, and safety-critical guidance because omissions may affect urgent decision-making. This study aimed to evaluate the clinical accuracy, completeness, readability, empathy, and reproducibility of responses generated by three LLMs to standardized caregiver questions about pediatric seizures.

Methods: This cross-sectional comparative study evaluated GPT-5.5, Gemini 3 Flash, and DeepSeek V3. Ten standardized caregiver questions addressing concerns about childhood seizures were submitted to each model in two runs during a 48-hour testing window in May 2026. Model identities were removed, and responses were evaluated by two pediatric neurologists using a 5-point rubric covering clinical accuracy and safety, completeness, readability and understandability, and empathy and tone. Inter-run consistency was assessed using Spearman's rank correlation coefficient.

Results: A total of 60 artificial intelligence-generated responses were evaluated. Expert-rated performance differed among the models in this exploratory analysis. Gemini received the highest overall scores within this prompt set and the May 2026 testing window, with a median score of 5.00; it was followed by GPT-5.5 and DeepSeek V3, which had median scores of 4.50 and 4.00, respectively. Although all models produced fluent and understandable responses, some outputs omitted clinically important safety information, including the five-minute emergency threshold. Inter-run consistency was moderate (Spearman's ρ=0.517). These findings are exploratory and may differ across model updates, testing periods, and clinical scenarios.

Conclusion: In this targeted expert evaluation of caregiver-oriented responses to pediatric seizures, model performance varied across LLMs, and repeated responses demonstrated only moderate consistency. Although Gemini achieved the highest scores within the prompt set and the testing window, the occasional omission of safety-critical advice, including the five-minute emergency threshold, indicates that LLMs should be used as supplementary sources of general information rather than as stand-alone tools for urgent seizure-related decisions. Relative model performance may change with model updates and when different clinical scenarios are evaluated.

Keywords: Pediatric seizures, large language models, ChatGPT, Gemini, DeepSeek, artificial intelligence, caregiver counseling, patient education

Introduction

Seizures are transient episodes of signs and/or symptoms resulting from abnormal excessive or synchronous neuronal activity in the brain (1). Febrile and non-febrile seizures are among the most common neurological emergencies in childhood. Because they often occur unexpectedly, they may cause considerable anxiety in parents, particularly after a first episode (2,3). Lacking immediate professional support, families are often driven by a sense of urgency to turn toward digital platforms for health-related guidance.

As large language model (LLM) applications become part of daily life, they are redefining how caregivers

Corresponding Author: Sema Bozkaya Yilmaz, MD, University of Health Sciences Türkiye, Bursa City Hospital, Clinic of Pediatric Neurology, Bursa, Türkiye
E-mail: semabozkayayilmaz@gmail.com **ORCID:** orcid.org/0000-0002-5389-5616
Received: 21.05.2026 **Accepted:** 06.08.2026 **Publication Date:** 24.09.2026

Cite this article as: Bozkaya Yilmaz S, Basarir G. Expert evaluation of large language models for caregiver counseling in pediatric seizures: a comparative study of ChatGPT, Gemini, and DeepSeek. Med Bull Haseki. 2026;64(4):229-236

navigate health-related information. While artificial intelligence (AI) technology has long been a staple in medicine, today's chat systems have become an increasingly accessible source of information for the general public. Recent studies and reviews have shown that LLMs are increasingly evaluated as tools for answering patient questions, generating patient education materials, and supporting caregiver-facing information; however, their performance remains variable, and concerns persist regarding accuracy, completeness, and safety (4-7). Built on distinct technical frameworks, ChatGPT, Gemini, and DeepSeek have emerged as some of the most accessible models for the general public today. Although LLMs have been increasingly studied in health care, their use in pediatric neurology remains less well characterized. Evidence is especially limited for caregiver-facing seizure counseling, where the quality of advice depends not only on accuracy but also on clear safety guidance (8).

Therefore, this study aimed to assess the clinical validity and communication quality of three widely used LLMs. Using standardized caregiver questions about pediatric seizures, we evaluated model-generated responses in terms of clinical accuracy, completeness, readability, empathy, and reproducibility. We hypothesized that the evaluated LLMs would differ in overall expert-rated performance and that repeated responses to identical caregiver prompts would show variability in clinically relevant content.

Materials and Methods

Compliance with Ethical Standards

This study did not involve human participants, animal subjects, patient-level data, private medical records, biological samples, identifiable personal data, patient images, or any clinical intervention. The prompts were standardized hypothetical caregiver questions developed by the authors, and the evaluated material consisted only of responses generated by publicly accessible LLMs. Therefore, institutional ethics committee approval and informed consent were not required. The authors declare no conflicts of interest and no financial support was received for this study.

Study Design and AI Models

This was a cross-sectional, prompt-based, comparative evaluation study designed to assess the quality and reproducibility of caregiver-oriented responses generated by LLMs in response to pediatric seizure-related questions. Three widely accessible LLMs were evaluated: ChatGPT/GPT-5.5 (OpenAI), Gemini 3 Flash (Google), and DeepSeek V3.

We used ten caregiver prompts, selected to represent common seizure-related questions in pediatric neurology practice. These prompts addressed acute seizure first aid, febrile seizures, emergency referral, electroencephalography interpretation, antiseizure treatment, neurodevelopmental concerns, the risk of later epilepsy, and daily-life precautions. Each question was entered twice into each of the three models, yielding 60 responses in total. The prompts were submitted through the publicly available web interfaces of the models without changing the default settings. To limit the effects of possible model updates, all queries were completed within a single 48-hour period in May 2026. Before the second run, previous conversations were cleared so that earlier prompts would not influence later responses.

Model names and versions were recorded as displayed in the public web interfaces at the time of testing. Because LLMs are frequently updated, the findings reflect model behavior during this specific testing period and may not be directly reproducible at later dates. The study workflow included prompt development, two independent model runs, collection and de-identification of responses, blinded expert evaluation, and statistical analysis, as summarized in Figure 1.

Prompt Development and Reference Standards

The prompts were developed by two pediatric neurology specialists to reflect common caregiver questions encountered in clinical practice. Rather than using a large number of automated prompts, the study deliberately focused on ten high-priority questions that would allow detailed expert assessment of clinically relevant and potentially safety-sensitive content. The full list of prompts is presented in Table 1.

Before formal scoring, the authors defined core expected content for seizure-related caregiver counseling based on established seizure terminology, febrile seizure guidance, and commonly accepted seizure first-aid principles. These predefined reference standards covered seizure first aid, emergency warning signs, febrile seizure counseling, diagnosis and treatment, developmental and epilepsy risks, and daily-life precautions. They were used to support the clinical accuracy, safety, and completeness domains of the scoring rubric. The predefined reference standards are summarized in Table 2.

Outcomes

The primary outcome was the overall expert-rated performance score assigned to each LLM-generated response. Secondary outcomes included domain-specific scores for clinical accuracy and safety; completeness; readability and understanding; and empathy and tone. Additional secondary outcomes included inter-run consistency of repeated model outputs, inter-rater agreement between the two evaluators, and presence of safety-critical omissions, including failure to mention the five-minute emergency threshold when clinically relevant.

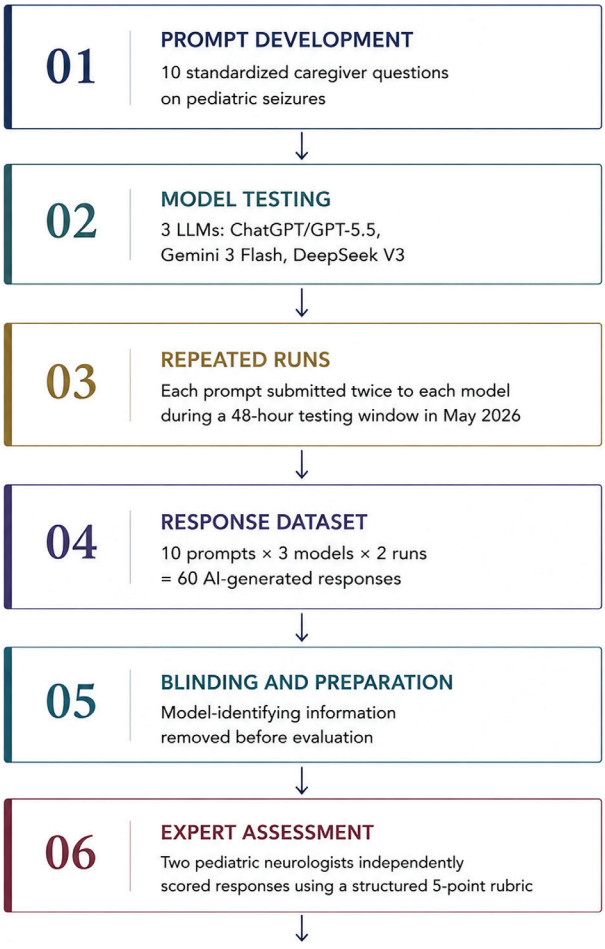

Figure 1. Study workflow. The workflow summarizes prompt development, model testing, repeated response generation, response de-identification, blinded expert evaluation, and statistical analysis. Ten standardized caregiver prompts were submitted twice to each of three large language models, resulting in 60 AI-generated responses
AI: Artificial intelligence

Evaluation Protocol

The responses generated by each model were collected in a standardized document. All information that could identify the source model was removed before evaluation to ensure that the assessors were blinded to model identity. However, complete blinding could not be guaranteed because model outputs may have recognizable stylistic characteristics.

The responses were independently evaluated by two pediatric neurologists. Before formal scoring, a pilot evaluation was performed using sample responses to align evaluators and ensure consistent interpretation of scoring domains. Each response was scored using a 5-point rubric ranging from 1, indicating very poor performance, to 5, indicating excellent performance. The rubric covered four domains: clinical accuracy and safety, completeness, readability and understanding, and empathy and tone. The scoring rubric is presented in Table 3. Rubric development followed a prespecified, stepwise process: the two pediatric neurologists first identified the clinically and communicatively essential domains; mapped the accuracy/safety and completeness domains to the predefined seizure reference standards; reviewed the readability and empathy domains in relation to dimensions commonly assessed in prior expert evaluations of patient- and caregiver-facing LLM outputs (5-7); and then pilot-tested the draft rubric on sample responses to harmonize domain interpretation and scoring thresholds before formal evaluation.

Each evaluator assigned four domain scores and a separate holistic overall performance score (1-5) for each response. The two evaluators' overall performance scores were averaged to yield the primary overall score for that response. Scores within each domain were likewise averaged across the two evaluators and reviewed descriptively. This overall score was used for descriptive model comparisons. The domain-specific scores were also reviewed to identify clinically relevant strengths and weaknesses in the generated responses.

Although the rubric was not a formally validated instrument, it was developed to reflect clinically relevant dimensions of caregiver-facing medical information, including clinical accuracy and safety, completeness, readability, and empathic communication. The four domains were defined a priori and selected because they correspond to recurring assessment dimensions in published evaluations of LLM-generated health information (5-7). The clinical accuracy, safety, and completeness domains were anchored to the predefined seizure-related reference standards described above, whereas readability and empathy were assessed as caregiver-oriented communication domains. Following the pilot review, the evaluators discussed discrepancies in domain interpretation and agreed on the operational meaning of each score level; the final scoring criteria and anchors are provided in Table 3. Therefore, the scoring process combined structured expert assessment with clinically anchored reference criteria.

Statistical Analysis

All data were analyzed using jamovi version 2.3 (The jamovi project, Sydney, Australia) and JASP version 0.18.3 (JASP Team, Amsterdam, the Netherlands). Descriptive statistics were reported as medians and interquartile ranges (IQR) because the score distributions were nonparametric and exhibited restricted ranges in some domains.

Table 1. Standardized caregiver prompts used for LLM evaluation

Prompt no.	Clinical topic	Standardized caregiver question
1	Acute seizure management	"My child had a seizure. What should I do?"
2	Febrile seizures-prognosis	"Are febrile seizures harmful?"
3	Febrile seizures-recurrence	"Can febrile seizures recur?"
4	Antiseizure medication	"Does a child who has had a seizure need medication?"
5	EEG interpretation	"Can seizures still occur if the EEG is normal?"
6	Emergency evaluation	"When should I take my child to the hospital?"
7	Seizure recognition	"How can I recognize a seizure in my child? What are the signs of seizures?"
8	Neurodevelopmental outcome	"Can seizures affect my child's intelligence or development?"
9	Epilepsy risk after febrile seizures	"Can febrile seizures develop into epilepsy later in life?"
10	Daily life precautions	"What precautions should be taken in the daily life of a child with seizures (sports, bathing, sleep)?"

EEG: Electroencephalography

Table 2. Predefined reference standards for seizure-related caregiver counseling

Domain	Expected core elements
Seizure first aid	Safe positioning, protection from injury, no restraint, nothing placed in the mouth, observation of breathing and seizure duration
Emergency warning signs	Urgent medical evaluation for seizures lasting approximately ≥5 minutes, recurrent seizures without recovery, breathing difficulty/cyanosis, prolonged altered consciousness, focal deficit, or first seizure
Febrile seizure counseling	Simple febrile seizures are usually benign; recurrence is possible; risk assessment depends on age, seizure features, neurological status, and family history
Diagnosis and treatment context	A normal EEG does not exclude seizures; antiseizure medication decisions depend on seizure type, recurrence risk, etiology, EEG findings, and specialist evaluation
Developmental and epilepsy risk	Long-term outcome depends on seizure type, underlying cause, seizure burden, and associated neurodevelopmental or neurological findings
Daily-life precautions	Supervision during bathing/swimming, avoidance of unsafe heights or high-risk activities, attention to sleep and medication adherence when relevant, and encouragement of age-appropriate activities

The reference standards were defined before formal scoring and were based on established seizure terminology, febrile seizure guidance, and commonly accepted seizure first-aid principles
EEG: Electroencephalography

Table 3. Scoring rubric for evaluating LLM-generated responses

Score	Clinical accuracy & safety	Completeness (adequacy)	Readability & understanding	Empathy & tone
5 (excellent)	Perfectly aligned with ILAE/AAP guidelines. No medical errors. Correct dosages/protocols.	All critical aspects (red flags, 5-min rule) covered comprehensively.	Professional yet perfectly accessible for laypersons. No jargon.	Highly supportive, reassuring, and culturally sensitive tone.
4 (good)	Accurate and safe advice. Only minor, non-critical details missing.	Covers all main points but lacks some secondary depth.	Clear and easy to read. Very few technical terms.	Reassuring and appropriate for a medical context.
3 (fair)	Generally safe but lacks specific clinical depth. Minor omissions in guidance.	Missing important secondary information (e.g., specific lifestyle tips).	Understandable but uses some medical jargon that may confuse parents.	Neutral tone. Helpful but lacks proactive empathy.
2 (poor)	Contains omissions that could lead to confusion or delayed action.	Significant gaps. Fails to mention several key clinical points.	Difficult to follow. Overly technical or excessively brief.	Clinical or dismissive tone. Fails to address parental anxiety.
1 (very poor)	Clinically unsafe. Incorrect dosages or dangerous advice.	Fails to answer the core question. Minimal or irrelevant content.	Unintelligible or completely inappropriate for a caregiver.	Apathetic, alarming, or inappropriate tone.

Scores ranged from 1 to 5, with higher scores indicating better performance
AAP: American Academy of Pediatrics, ILAE: International League Against Epilepsy, LLM: large language model

The Kruskal-Wallis H test was used for exploratory comparison of overall performance scores among the three models, with post-hoc comparisons performed using the Dwass-Steel-Critchlow-Fligner (DSCF) procedure. Because the same standardized prompts were repeated across models and runs, the observations were not fully independent. A mixed-effects model or a generalized estimating equations approach would ordinarily be preferable for this repeated-prompt structure. However, these models were not fitted because the dataset included only ten prompt clusters, two runs per model, and strongly restricted score distributions with a ceiling effect, conditions that could yield unstable or poorly identified estimates. Therefore, statistical comparisons were interpreted as exploratory and considered together with descriptive score distributions and clinically relevant response patterns rather than as confirmatory evidence. Accordingly, the reported p-values are only hypothesis-generating and should not be interpreted as confirmatory evidence of superiority between models.

Inter-run consistency was assessed by correlating the mean overall scores from Run 1 and Run 2 across 30 matched prompt-model pairs using Spearman's rank correlation coefficient (ρ); exact agreement was also reported descriptively. Inter-rater agreement was primarily assessed using absolute percentage agreement. Because score distributions were restricted and skewed in some qualitative domains, particularly empathy and completeness, chance-corrected agreement statistics were difficult to interpret. Therefore, absolute percent agreement was reported as a descriptive measure of inter-rater agreement and was not considered a substitute for chance-corrected reliability statistics.

To further describe clinical consensus, scores were also interpreted using a dichotomous framework: scores of 3-5 were considered clinically acceptable, whereas scores of 1-2 were considered unacceptable. A p-value below 0.05 was used as a threshold for statistical significance, but all inferential findings were interpreted cautiously in light of the exploratory design and repeated-prompt structure. No causal, confirmatory, or broadly generalizable inferences were made from these analyses.

Results

Comparative Performance of AI Models

A total of 60 AI-generated responses were evaluated, with 20 responses obtained from each model. These responses were produced by submitting ten standardized prompts to each model in two independent runs. In the exploratory comparison, overall performance scores differed among the three models (Kruskal-Wallis $\chi^2(2)=54.6$, $p<0.001$). Gemini had the most favorable overall ratings in this prompt set, with scores clustering at the maximum value (median 5.00, IQR: 5.00-5.00). GPT-5.5 had a median of 4.50 (IQR: 4.50-4.50), whereas DeepSeek V3 had lower overall scores, with a median of 4.00 (IQR: 4.00-4.00). The distribution of overall performance scores across the three models is shown in Figure 2. Within this prompt set, Gemini received the maximum score for all evaluated responses, indicating a ceiling effect in overall performance scores. This pattern describes only the tested prompts and the May 2026 model versions and should not be interpreted as stable superiority across updates or other clinical contexts.

Post-hoc Pairwise Comparisons

In exploratory post-hoc comparisons using the DSCF procedure, Gemini received higher overall performance scores than GPT-5.5 ($W=8.73$, $p<0.001$) and DeepSeek V3 ($W=8.73$, $p<0.001$), while GPT-5.5 received higher scores than DeepSeek V3 ($W=7.44$, $p<0.001$). Qualitatively, GPT-5.5 responses were generally easy for caregivers to read and understand, whereas some responses contained less clinical detail than Gemini responses; these domain-level observations were descriptive and were not tested statistically. Which focused on emergency evaluation: GPT-5.5 used an authoritative tone but did not explicitly mention the five-minute threshold for emergency evaluation; this omission is clinically relevant because prolonged seizures require urgent medical assessment.

Inter-rater Reliability Analysis

Overall, a high level of clinical consensus was observed between the two blinded evaluators. The restricted score range-particularly because one evaluator frequently

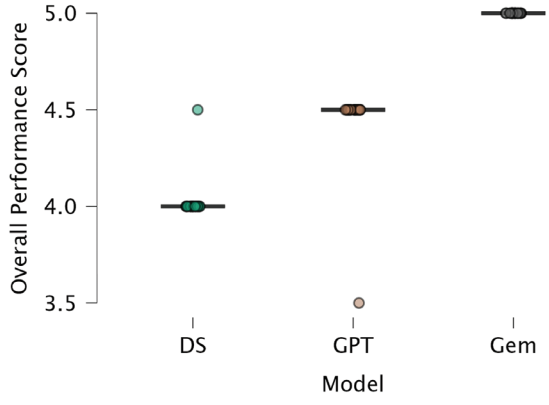

Figure 2. Distribution of overall performance scores across AI models. The plot illustrates the distribution of expert-rated overall performance scores for DeepSeek, GPT-5.5, and Gemini. Individual data points and boxplots are shown for each model. Gemini scores clustered at the maximum value, indicating a ceiling effect within the present prompt set
AI: Artificial intelligence

assigned maximum scores in some domains-limited the usefulness of chance-corrected agreement statistics. Because of the restricted score range and lack of variance in these categories, a formal Cohen's kappa coefficient could not be meaningfully calculated. Therefore, absolute percent agreement was reported descriptively. In the completeness domain, the absolute percent agreement was 66.7%. In the empathy domain, evaluators appeared to apply different subjective thresholds, with scores of 3-5 versus 4-5; the exact-match agreement was 33.3%. Although exact agreement for empathy was lower, the discrepancy reflected differences in subjective communication thresholds rather than a major disagreement regarding clinical safety. When scores were dichotomized into clinically acceptable and unacceptable categories, inter-rater agreement was 100%. This suggests that the lower exact-match rates mainly reflect differences in scoring thresholds for communication quality, rather than disagreement about clinical safety. Despite variability in exact scores, no response from any model was rated as clinically poor (1) or unsafe (2) by either evaluator in any evaluated category.

Reliability and Reproducibility Analysis

Across the 30 matched prompt-model pairs, mean overall scores were strongly correlated between the first and second runs (Spearman's ρ=0.922, $p<0.001$), with exact agreement in 28 pairs (93.3%). Exact agreement was observed in 9/10 GPT-5.5 pairs, 10/10 Gemini pairs, and 9/10 DeepSeek V3 pairs. Nevertheless, repeated responses were not textually identical, and qualitative differences in detail and emphasis were observed across runs. This variability is relevant because caregivers may receive different safety information depending on when and how the same question is asked. All generated responses were successfully recorded and included in the analysis.

Discussion

Seizures are among the most frequent neurological emergencies in children, representing nearly 3% of pediatric emergency visits and affecting about 5% of children worldwide (9,10). For many parents, witnessing a seizure is intensely frightening, and some describe it as a moment in which they fear losing their child (11,12). In moments of uncertainty, families may seek quick, accessible guidance before they can reach a clinician. With conversational AI tools now widely available, caregivers may also turn to LLM-based systems for explanations, reassurance, or practical advice. For this reason, incomplete or inaccurate AI-generated advice in pediatric seizure counseling requires careful evaluation.

In the present exploratory evaluation, overall expert-rated performance differed among the evaluated AI models. For this specific set of prompts, Gemini received higher overall performance scores than GPT-5.5 and DeepSeek V3. The variability in performance across the evaluated models is consistent with recent systematic reviews showing that LLMs may provide useful patient-facing medical information but remain limited by inconsistent performance, incomplete responses, accuracy gaps, and safety concerns. Qualitatively, Gemini's responses more consistently included guideline-concordant elements, including International League Against Epilepsy-based seizure terminology and American Academy of Pediatrics-aligned recommendations for febrile seizure management; however, this was a descriptive observation and was not tested statistically. In addition, the overall performance result should be interpreted within the limitations of the selected prompt set and the testing window. The ceiling effect observed with Gemini's uniform maximum scores may also indicate that the 5-point rubric was not sensitive enough to distinguish between strong responses at the upper end of performance. Thus, Gemini's higher ratings should be understood as a context-specific finding from this prompt set and testing period rather than evidence of general or enduring model superiority; rankings may change after model updates or when different pediatric seizure scenarios are tested.

Although all three models were able to generate fluent and understandable responses, our findings show that linguistic fluency should not be assumed to reflect clinical reliability. One notable observation was that responses could sound fluent and authoritative while still missing clinically relevant content-a pattern we describe as a potential "confidence gap". For example, GPT-5.5's response to Prompt 6 was clearly written and generally helpful but did not explicitly mention the five-minute threshold for emergency evaluation. This omission is clinically relevant because seizures lasting five minutes or longer are generally treated as prolonged seizures and require urgent medical evaluation. Although no response was rated as manifestly unsafe, such omissions may nevertheless be important in seizure counseling for caregivers, where urgent warning signs must be recognized promptly. For caregivers without medical training, a confidently written answer may appear trustworthy even when it is incomplete.

The reproducibility analysis added another important dimension to clinical reliability. The pooled Spearman correlation of 0.922 indicated strong agreement between run-level overall scores. However, because score ranges were highly restricted within models and the pooled correlation was partly influenced by between-model differences, this finding should not be interpreted as

complete content-level reproducibility. Repeated responses were not textually identical and differed in detail and emphasis; such variation may affect the communication of safety warnings, emergency thresholds, and follow-up recommendations.

Differences in scoring for subjective domains, such as empathy, also highlight the difficulty of standardizing the quality of communication. What one evaluator considers sufficiently reassuring may be judged as less supportive by another. Similar concerns were reported in caregiver-oriented LLM evaluations, where models produced clear and accessible responses but still omitted relevant practical details when compared with professional reference standards (6). Future studies should include caregivers directly, since expert-rated quality may not fully reflect how families understand, trust, or act upon AI-generated advice in real-life settings. It will also be important to determine whether these responses reduce anxiety or instead increase confusion, false reassurance, or unnecessary healthcare-seeking behavior.

Study Limitations

Several limitations should be acknowledged. The study was based on ten standardized clinical scenarios rather than real-life caregiver interactions. This focused prompt set enabled expert neurologists to examine high-stakes, seizure-related questions in detail; however, the findings should not be generalized to all pediatric seizure-related situations. Specifically, the study was limited to ten standardized prompts, two expert evaluators, and a single 48-hour testing window in May 2026. Consequently, the results may not represent the full range of caregiver wording, seizure types, levels of clinical urgency, sociocultural contexts, or longitudinal changes in model behavior. Because LLMs are updated frequently, these findings should be considered a snapshot of model performance during the specific testing period and may not be generalizable to future model versions. In addition, because the dataset was based on repeated standardized prompts, the observations were not fully independent; therefore, the statistical comparisons should be interpreted as exploratory rather than confirmatory.

Although explicit model identifiers were removed prior to evaluation, complete blinding could not be guaranteed, as some models may exhibit recognizable stylistic response patterns. The responses were evaluated by expert pediatric neurologists using a structured rubric; however, the rubric was not a formally validated instrument. In addition, the study did not directly assess caregivers' perceptions, comprehension, trust, anxiety, healthcare-seeking behavior, or real-life decision-making after receiving AI-generated advice. Therefore, expert-rated quality may not fully reflect how caregivers would understand or act upon these responses during an actual seizure-related situation. Finally, the study evaluated text-based responses only and did not examine voice-based, multimodal, or interactive use of LLMs in real-world settings. Despite these limitations, the study has several strengths, including the use of standardized clinically relevant prompts, predefined reference standards, blinded independent evaluation by two pediatric neurologists, and repeated model testing to assess reproducibility. These methodological features provide a structured and clinically grounded comparison of LLM performance in caregiver-oriented pediatric seizure counseling.

Conclusion

In this targeted expert evaluation of 60 LLM-generated responses to caregiver questions about pediatric seizures, Gemini achieved the highest scores within the tested prompt set. However, overall scores were highly consistent across runs, although repeated responses differed in detail and emphasis, and some responses omitted clinically important emergency advice, including the five-minute emergency threshold. This comparative finding is limited to the May 2026 testing window and to the clinical scenarios examined and may change with model updates or alternative prompt sets. These findings suggest that LLMs may support general caregiver education, but they should not be used as stand-alone sources for urgent seizure-related decisions. Future studies should examine how caregivers understand, trust, and act upon AI-generated seizure advice in real-life settings.

Ethics

Ethics Committee Approval: This study did not involve human participants, animal subjects, or the use of private medical records. As the evaluation was conducted using publicly available large language models and standardized clinical prompts, institutional review board approval was not required.

Informed Consent: Informed consent was not required, as the study did not involve human participants or the collection or use of identifiable patient-level data.

Footnotes

Authorship Contributions

Concept: S.B.Y., G.B., Design: S.B.Y., G.B., Data Collection or Processing: S.B.Y., G.B., Analysis or Interpretation: S.B.Y., G.B., Literature Search: S.B.Y., G.B., Writing: S.B.Y., G.B.

Conflict of Interest: The authors declared that there were no conflicts of interest.

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

Declaration Regarding the Use of AI and AI-Assisted Technologies: This study evaluates responses

generated by GPT-5.5, Gemini 3 Flash, and DeepSeek V3. The study design, prompt selection, scoring framework, clinical evaluation, statistical analysis, interpretation of the findings, and final manuscript decisions were carried out by the authors. AI tools did not replace author judgment at any stage. The authors reviewed and approved the final manuscript and take full responsibility for its content.

References

1. Fisher RS, van Emde Boas W, Blume W, et al. Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia. 2005;46:470-2.
2. Wirrell E, Turner T. Parental anxiety and family disruption following a first febrile seizure in childhood. Paediatr Child Health. 2001;6:139-43.
3. Save-Pédebos J, Bellavoine V, Goujon Eet al. Difference in anxiety symptoms between children and their parents facing a first seizure or epilepsy. Epilepsy Behav. 2014;31:97-101.
4. Yun HS, Bickmore T. Online health information-seeking in the era of large language models: cross-sectional web-based survey study. J Med Internet Res. 2025;27:e68560.
5. Jacobs MMG, Oosterhoff JHF, Agricola R, van der Weegen W. Large language models versus healthcare professionals in providing medical information to patient questions: a systematic review. Int J Med Inform. 2026;209:106250.
6. Pérez-Esteve C, Guilabert M, Matarredona V, et al. AI in home care-evaluation of large language models for future training of informal caregivers: observational comparative case study. J Med Internet Res. 2025;27:e70703.
7. Khosravi M, Zamaninasab Z, Mojtabaeian SM, Dindar Demiray EK, Arab-Zozani M. A systematic review of the limitations of large language models in generating healthcare content. PLOS Digit Health. 2026;5:e0001354.
8. Fennig U, Yom-Tov E, Savitzky L, et al. Bridging the conversational gap in epilepsy: using large language models to reveal insights into patient behavior and concerns from online discussions. Epilepsia. 2025;66:686-99.
9. Smith RA, Martland T, Lowry MF. Children with seizures presenting to accident and emergency. Emerg Med J. 1996;13:54-8.
10. Subcommittee on Febrile Seizures; American Academy of Pediatrics. Neurodiagnostic evaluation of the child with a simple febrile seizure. Pediatrics. 2011;127:389-94.
11. Hall-Parkinson D, Tapper J, Melbourne-Chambers R. Parent and caregiver knowledge, beliefs, and responses to convulsive seizures in children in Kingston, Jamaica: a hospital-based survey. Epilepsy Behav. 2015;51:306-11
12. Besag FM, Nomayo A, Pool F. The reactions of parents who think that a child is dying in a seizure–in their own words. Epilepsy Behav. 2005;7:517-23.