

Safety, Accuracy, Empathy, Information Quality, and Readability of Large Language Models in Answering Caregiver Questions About Pediatric Burn First Aid and Inpatient Care: A Cross-Sectional Comparative Study

Chun Wang

Beijing Children's Hospital Capital Medical University

Ying Li

Beijing Children's Hospital Capital Medical University

Yihan Ma

Beijing Children's Hospital Capital Medical University

Zhishui Wu²

Capital Medical University, Third Affiliated Hospital of Soochow University

Qin Yang¹

yangqin@bch.com.cn

Beijing Children's Hospital Capital Medical University

Research Article

Keywords: Large language models, Artificial intelligence, Pediatric burns, Burn first aid, Caregiver education, Health information quality, Readability

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10862831/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Pediatric burn injuries are common household injuries, particularly among young children, and hot-liquid scalds are a major cause of pediatric burn morbidity. The quality of burn first aid provided by caregivers can affect pain, burn wound depth, wound healing, the need for surgical intervention, and scarring. As large language model (LLM)-driven chatbots become widely available sources of health information, caregivers may consult these tools in time-critical situations requiring burn first aid; however, their performance in pediatric burn scenarios remains insufficiently characterized.

Objective

To compare the safety, accuracy, empathy, information quality, and readability of four widely accessible LLM chatbots in answering caregiver-facing questions about pediatric burn first aid and inpatient care.

Methods

In this cross-sectional comparative evaluation, a 36-item question set was developed from an initial pool of 66 caregiver-facing questions derived from clinical guidelines, consensus statements, official patient-education materials, systematic reviews, and common pediatric burn care concerns. The same questions were entered independently into ChatGPT, Gemini, DeepSeek, and Doubao, yielding 144 chatbot responses. Five blinded raters evaluated outputs using predefined safety and accuracy criteria, an empathy scale, DISCERN, EQUIP, JAMA benchmark criteria, and the Global Quality Score. Readability was assessed using the Automated Readability Index, Coleman-Liau Index, Flesch-Kincaid Grade Level, Gunning Fog Index, Simple Measure of Gobbledygook, and Flesch Reading Ease Score. Inter-rater agreement and paired question-level comparisons were examined.

Results

Inter-rater agreement was high for safety assessment (Fleiss $\kappa = 0.860$; 95% CI, 0.784–0.924; $P < 0.001$) and for the other rater-scored outcomes (ICC range, 0.831–0.899; all $P < 0.001$). Responses were classified as safe in 33 of 36 cases for ChatGPT (91.7%), 31 of 36 for Gemini (86.1%), 30 of 36 for DeepSeek (83.3%), and 29 of 36 for Doubao (80.6%). Accuracy scores differed across models ($P = 0.002$), although the overall effect size was small (Kendall's $W = 0.137$). Empathy scores showed a larger between-model difference ($P < 0.001$; Kendall's $W = 0.564$), with Doubao receiving lower ratings than the other models. DISCERN, EQUIP, and Global Quality Scores differed across models (all $P < 0.001$), whereas JAMA benchmark scores did not ($P = 0.330$). All readability indices differed across models (all $P <$

0.001). Gemini and DeepSeek generally produced responses with lower estimated grade levels than ChatGPT and Doubao.

Conclusions

The evaluated LLM chatbots generally received favorable safety and accuracy ratings when answering caregiver questions about pediatric burns. However, potentially unsafe responses, limited source transparency, variable empathy, and substantial readability barriers remained. Future pediatric burn information systems should prioritize concise, action-oriented guidance, plain-language communication, transparent evidence attribution, and clear indications for urgent medical assessment.

Clinical trial number:

Not applicable.

Introduction

Pediatric burns remain a major and preventable cause of morbidity worldwide [1–3]. The World Health Organization estimates that burns account for approximately 180,000 deaths annually, with a disproportionate burden in low- and middle-income countries [1]. Nonfatal burns may result in prolonged hospitalization, disfigurement, functional impairment, psychosocial burden, and substantial family stress [2, 3]. Children are particularly vulnerable because burn mechanisms vary by developmental stage, and children's skin is thinner than adult skin, increasing the risk of a deeper burn injury after comparable thermal exposure [4]. These characteristics make timely burn first aid, early recognition of high-risk burn injuries, and caregiver education essential components of pediatric burn care [5–10].

Caregivers usually provide the initial care immediately after a burn, before professional assessment is available. Guideline-concordant burn first aid emphasizes stopping the burning process; removing clothing and jewelry near the burn when safe, while leaving adherent materials in place; cooling the burn wound with cool running tap water; loosely covering it with a clean, non-adherent dressing or plastic cling film; avoiding ice and non-recommended topical agents; and obtaining urgent medical assessment for high-risk injuries [5–10]. After hospital admission or specialist burn assessment, caregivers continue to require practical guidance on burn wound care, dressing changes, pain management, signs of burn wound infection, nutritional support, post-burn pruritus, scar management, rehabilitation, and return to age-appropriate activities [5, 8, 11]. These questions require not only factual accuracy but also clear escalation advice, appropriate emotional support, and language that is accessible to non-specialists [29, 30].

LLM chatbots have rapidly become accessible sources of health information. Their conversational interfaces, immediate availability, and ability to generate apparently tailored explanations may make

them appealing to families seeking advice outside routine clinical encounters [12–14]. Previous studies have reported that chatbot responses may be perceived as high quality or empathetic in selected settings involving patient questions [13]. At the same time, medical accuracy, consistency, source transparency, and safety remain variable, particularly in patient-facing contexts [15–18]. These limitations are consequential in time-sensitive pediatric conditions, in which delayed referral, inappropriate home treatment, or failure to identify red flags may cause harm [5–8, 15–18].

Caregiver questions about pediatric burns provide an important context for evaluating LLM performance because they encompass burn first aid, inpatient nursing care, burn wound care, rehabilitation, and family-centered communication [5, 8, 11]. A fluent or reassuring response may still be unsafe if it recommends inappropriate burn cooling, understates the risks associated with chemical or electrical injuries, overlooks suspected inhalation injury or possible non-accidental injury, endorses non-recommended topical agents or home remedies, or fails to direct caregivers to urgent medical assessment when clinically indicated [5–8]. Conversely, technically accurate information may have limited practical value if distressed caregivers cannot quickly understand and act on it [29, 30].

Evidence remains limited regarding the performance of widely used LLM chatbots in pediatric burn scenarios from the caregiver perspective. Existing evaluations have often focused on general medical questions, examination-style tasks, adult conditions, or single-model assessments, and many have not assessed safety, accuracy, empathy, information quality, and readability simultaneously [17–19]. We therefore performed a multidimensional, guideline-referenced comparative evaluation of ChatGPT, Gemini, DeepSeek, and Doubao in response to caregiver-facing questions on pediatric burn first aid and inpatient care.

Pediatric burn scenarios present a uniquely high-stakes context for evaluating LLMs because the initial response to a burn injury is often time-critical, the quality of first aid directly affects clinical outcomes, and distressed caregivers may be particularly vulnerable to fluent but unsafe or incomplete information. Furthermore, the combination of acute first-aid decisions and ongoing inpatient care questions creates a broad spectrum of content that tests models across multiple domains simultaneously—from emergency triage to wound care to psychosocial support. Evaluating LLMs in this specific context therefore provides insight into both their potential utility and their limitations in pediatric urgent-care and family-education settings.

Methods

Study Design and Reporting Framework

This cross-sectional comparative study evaluated LLM chatbots in response to caregiver-facing questions related to pediatric burn first aid and inpatient care. The study was designed and reported in accordance with the Chatbot Assessment Reporting Tool (CHART), which provides reporting recommendations for studies evaluating generative AI chatbots that summarize clinical evidence or provide health advice [18,19].

Four publicly accessible, widely used LLM chatbots were assessed: ChatGPT, Gemini, DeepSeek, and Doubao. Each model answered the same question set, and outputs were evaluated across five domains: safety, accuracy, empathy, information quality, and readability. The study used hypothetical questions and publicly accessible chatbot-generated responses; no real patients, clinical encounters, identifiable personal information, or protected health information were involved. The overall study workflow is shown in Figure 1.

Development of the Caregiver-Facing Question Set

The question set was developed to represent common information needs among caregivers of children with burns. Two investigators independently collected candidate questions from publicly available sources, including clinical guidelines, consensus statements, official patient-education materials, systematic reviews, and frequently discussed caregiver concerns in pediatric burn care [5-11]. Candidate questions were phrased in language likely to be used by parents or other family caregivers rather than in clinician-oriented terminology.

An initial pool of 66 questions was generated. Two investigators independently screened the questions according to predefined eligibility criteria. Questions were included if they were caregiver-facing, relevant to pediatric burn first aid or inpatient care, answerable using guideline- or consensus-based evidence, and suitable for evaluating LLM-generated health education. Questions were excluded if they were duplicates or near-duplicates, overly technical, administrative or cost-related, product- or brand-specific, dependent on local circumstances, outside the pediatric burn context, or if they required individualized diagnosis or treatment beyond general health education. Disagreements were resolved through discussion, with senior clinical adjudication when necessary. Thirty-six questions were retained for the final evaluation. The screening record is provided in Supplementary Material 1.

To further ensure that the final question set reflected real-world caregiver information needs, candidate items were cross-referenced with common concerns identified from the patient and family education sections of the cited guidelines, official patient-information sources, and systematic reviews. Questions were also phrased in plain, non-technical language that matched the typical wording used by parents and family caregivers in clinical and online health-inquiry settings.

Reference-Answer Mapping Table

Before model evaluation, a structured reference-answer mapping table was developed to support safety and accuracy scoring. Two investigators independently identified key evidence-based answer elements from pediatric burn first aid and acute management guidelines, consensus statements, official burn care recommendations, and systematic reviews [5-11]. For each retained question, the mapping table specified the essential content, required safety warnings, and clinically important caveats expected in an appropriate response.

The two investigators reviewed and reconciled the mapping table. They resolved differences in wording or interpretation through discussion with a senior clinician experienced in pediatric burn care. The final mapping table drew on more than 15 evidence sources and served as the reference standard for evaluating accuracy and safety. The complete mapping table is provided in Supplementary Material 2.

Model Selection and Model Information

The four models were selected because they were among the most accessible general-purpose LLM chatbots available to caregivers at the time of data collection. The selection included both international and Chinese chatbot platforms to reflect tools that potential users may access when seeking pediatric burn information. Model identifiers, providers, access routes, source status, query dates, and query location are summarized in Table 1. When an interface or provider documentation displayed an exact model version, it was recorded. When a specific model version was not displayed, the product name and query date were recorded.

Table 1. Model characteristics and query conditions

Model	Provider	Access route	Version and source status	Query date and location
ChatGPT	OpenAI	Web interface	GPT-5.5 Instant	July 1, 2026; Beijing, China
Gemini	Google	Web interface	Gemini 3.5 Flash	July 1, 2026; Beijing, China
DeepSeek	DeepSeek	Web interface	DeepSeek-V4 Preview	July 1, 2026; Beijing, China
Doubao	ByteDance	Web interface	Doubao-Seed-2.0	July 1, 2026; Beijing, China

Query Procedure and Response Collection

For each model, queries were submitted using the default generation parameters available in the web interface at the time of querying. Specific numerical values for parameters such as temperature, top-p, or top-k were not adjustable or displayed by the interface and therefore were not recorded. No system prompts, role assignments, few-shot examples, or corrective feedback were used during the query process. Each question was submitted in a new chat session to minimize carryover effects from previous exchanges.

Responses were collected during a predefined query period from the specified location. The full date of querying was recorded for each chatbot. Outputs were copied verbatim immediately after generation and stored in a standardized extraction file. Non-content formatting, such as interface spacing and decorative elements, was removed during data cleaning. The substantive content of the responses was not edited, paraphrased, or corrected. Each model generated one response per question, yielding 144 responses in total. The cleaned question-and-answer dataset is provided in Supplementary Material 3.

Evaluation Instruments

Safety was assessed as a binary outcome. Responses were classified as potentially unsafe if they contained advice that could plausibly cause harm, delay urgent medical assessment, recommend inappropriate home management, omit critical red-flag information, or conflict with established pediatric burn care recommendations [5-8]. Safety was coded as 0 for safe and 1 for potentially unsafe. For descriptive model-level safety proportions, the final response-level classification was determined by majority vote across the five raters.

Accuracy was rated on a 5-point ordinal scale according to concordance with the predefined reference-answer mapping table. A score of 1 indicated a completely inaccurate response that contradicted the reference standard or contained major misinformation. A score of 2 indicated mostly inaccurate content with minimal correct information and substantial omissions or errors. A score of 3 indicated partial accuracy with correct information but notable omissions, ambiguity, or minor inaccuracies. A score of 4 indicated a mostly accurate, guideline-concordant response with only minor omissions or imprecision. A score of 5 indicated a comprehensive response fully consistent with the reference standard and free of clinically relevant omissions.

Empathy was rated on a 5-point scale reflecting whether responses acknowledged caregiver concern, used supportive and respectful language, avoided blame, and communicated in a manner appropriate for parents or family members of pediatric patients.

Information quality was evaluated using DISCERN [20,21], EQIP [22], JAMA benchmark criteria [23], and the Global Quality Score (GQS) [24]. DISCERN was used to assess the reliability and quality of treatment-related consumer health information. All 16 items were summed to generate a total score from 16 to 80, with higher values indicating better quality. EQIP was used to assess the quality and usability of patient-facing health education materials. JAMA benchmark criteria assessed authorship, attribution, disclosure, and currency. GQS provided an overall judgment of usefulness, structure, and flow for patient-facing health information.

Readability was evaluated using the Automated Readability Index (ARI), Coleman-Liau Index (CLI), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES) [25-29]. For ARI, CLI, FKGL, GFI, and SMOG, higher scores indicate greater reading difficulty; for FRES, higher scores indicate easier readability. Readability findings were interpreted in relation to commonly recommended readability standards for patient-facing health materials [30,31].

Raters, Training, and Blinding

Five independent raters evaluated the model outputs. The rater panel comprised professionals in pediatric burn care, pediatric surgery, pediatric nursing, or health education; individual clinical or research experience ranged from 11 to 33 years. Before formal scoring, the raters received a standardized manual defining each evaluation metric, scoring rule, and example response. A calibration session was conducted using pilot responses that were not included in the final analysis.

For formal scoring, responses were anonymized and randomly ordered. Model names and interface-identifying information were removed. Raters were blinded to the chatbot that generated each response. Each rater independently scored all 144 responses. Individual ratings were retained for the primary analysis and inter-rater agreement assessment; no consensus scores were substituted for the original ratings.

Statistical Analysis

Descriptive statistics were used to summarize performance across all domains. Categorical variables were reported as frequencies and percentages. Continuous or ordinal outcomes were summarized as means with standard deviations or medians with interquartile ranges, depending on distribution and metric type.

Inter-rater agreement was assessed before comparative analyses. Fleiss' kappa was used for binary safety ratings [32]. For accuracy, empathy, DISCERN, EQIP, JAMA, and GQS, reliability was estimated using an intraclass correlation coefficient based on a two-way random-effects, absolute-agreement, single-measure model, denoted ICC(2,1) [33]. Estimates were reported with 95% confidence intervals.

Because the same questions were answered by all models, paired question-level analyses were used. Overall differences in safety were assessed using Cochran's Q test [34], with pairwise McNemar tests when appropriate [35]. Overall differences in accuracy, empathy, information quality scores, and readability indices were assessed using Friedman tests [36]. Kendall's W was calculated as a measure of effect size for overall inter-model differences [37], with 95% confidence intervals estimated by nonparametric bootstrap resampling by question. Pairwise comparisons for non-binary outcomes used paired Wilcoxon signed-rank tests [38], with Benjamini-Hochberg adjustment for multiple comparisons [39]. A two-sided P value < 0.05 was considered statistically significant. Analyses were performed in R [40] using RStudio version 2026.06.0+242 (Posit Software, PBC, Boston, MA, USA).

Ethical Considerations and Data Availability

This study evaluated hypothetical caregiver-facing questions and publicly accessible chatbot outputs. No patients, clinical encounters, identifiable personal information, or protected health information were involved. Raters assessed de-identified chatbot responses in their professional capacity. Formal institutional ethics review was therefore not required under local policy.

The question-screening record, reference-answer mapping table, and cleaned question-and-answer dataset are available in the Supplementary Materials. The scoring dataset and statistical code are available from the corresponding author upon reasonable request.

Results

Inter-rater Agreement

A total of 144 complete chatbot responses were evaluated by five raters. Inter-rater agreement was high for safety assessment (Fleiss' kappa = 0.860, 95% CI 0.784–0.924; $P < 0.001$). For the remaining metrics, ICC(2,1) estimates ranged from 0.831 for empathy to 0.899 for DISCERN, with all confidence intervals excluding zero and all P values < 0.001 . Specifically, ICC estimates were 0.857 for accuracy, 0.831 for empathy, 0.899 for DISCERN, 0.850 for EQIP, 0.843 for JAMA, and 0.843 for GQS.

Safety, Accuracy, and Empathy

Safety, accuracy, and empathy results are summarized in Figure 2 and Table 2. Safe responses constituted the majority of outputs for all models, although potentially unsafe outputs occurred with every model. The observed proportions classified as safe were 33/36 (91.7%) for ChatGPT, 31/36 (86.1%) for Gemini, 30/36 (83.3%) for DeepSeek, and 29/36 (80.6%) for Doubao. Most potentially unsafe classifications reflected incomplete escalation advice, insufficient emphasis on red flags, or inadequate warnings against contraindicated home practices; detailed qualitative classifications are provided in Supplementary Material 5.

Accuracy scores differed significantly across models (overall $P = 0.002$), with a small overall effect size (Kendall's $W = 0.137$, 95% CI 0.044–0.304). The median accuracy score was 4.00 for all four models, indicating generally favorable factual concordance with guideline-based reference answers. However, the statistically significant difference and the distribution patterns suggest that clinically meaningful variation emerged in the completeness of red-flag communication, the precision of escalation advice, and the prominence of safety warnings—dimensions that are not captured by median scores alone. This underscores the importance of multidimensional assessment beyond simple factual accuracy.

Table 2. Accuracy and empathy scores across models.

Metric	ChatGPT	Gemini	DeepSeek	Doubao	Overall P	Kendall's W (95% CI)
Accuracy	4.00 [4.00, 5.00]	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [3.80, 4.00]	0.002	0.137 (0.044–0.304)
Empathy	4.00 [4.00, 4.00]	4.00 [4.00, 5.00]	4.00 [3.80, 4.00]	3.00 [3.00, 4.00]	<0.001	0.564 (0.469–0.694)

Values are presented as median [Q1, Q3] unless otherwise indicated. Overall P values were obtained using Friedman tests.

Reliability and Information Quality

Reliability and information quality outcomes are presented in Figure 3 and Table 3. Significant model differences were observed for DISCERN, EQIP, and GQS (all $P < 0.001$). ChatGPT and Gemini had higher median DISCERN and EQIP scores than DeepSeek and Doubao. GQS also differed significantly across models, with lower scores most evident for Doubao. In contrast, JAMA scores were similar across models and did not differ significantly ($P = 0.330$). The uniformly low JAMA scores suggested limited

reporting of authorship, attribution, disclosure, and currency across all four models. Pairwise comparisons are provided in Supplementary Material 4.

Table 3. Reliability and information quality scores across models.

Metric	ChatGPT	Gemini	DeepSeek	Doubao	Overall P	Kendall's W (95% CI)
DISCERN	73.50 [73.20, 74.00]	73.40 [72.75, 74.00]	70.80 [69.60, 72.25]	70.10 [68.00, 70.80]	<0.001	0.596 (0.470–0.738)
EQIP	85.00 [83.00, 86.00]	85.00 [82.75, 86.00]	79.50 [75.50, 81.00]	75.50 [73.00, 80.00]	<0.001	0.496 (0.375–0.645)
JAMA	2.00 [2.00, 2.00]	2.00 [2.00, 2.00]	2.00 [2.00, 2.00]	2.00 [2.00, 2.00]	0.330	0.032 (0.009–0.142)
GQS	4.00 [4.00, 4.00]	4.00 [4.00, 5.00]	4.00 [3.80, 4.00]	3.80 [3.00, 4.00]	<0.001	0.228 (0.113–0.404)

Values are presented as median [Q1, Q3] unless otherwise indicated. Overall P values were obtained using Friedman tests.

Readability

SMOG means (Figure 4; Table 4) were 15.30 ± 3.75 for ChatGPT, 12.42 ± 1.84 for Gemini, 12.11 ± 1.58 for DeepSeek, and 15.59 ± 3.25 for Doubao (*P* < 0.001; Kendall's *W* = 0.443, 95% CI 0.268–0.640). FRES means were 35.68 ± 23.96 for ChatGPT, 51.38 ± 12.09 for Gemini, 56.70 ± 10.28 for DeepSeek, and 31.24 ± 19.00 for Doubao (*P* < 0.001; Kendall's *W* = 0.502, 95% CI 0.351–0.689). Pairwise results: Supplementary Material 4.

Table 4. Readability scores across models.

Metric	ChatGPT	Gemini	DeepSeek	Doubao	Overall P	Kendall's W (95% CI)
ARI	19.26 ± 11.36	10.65 ± 2.63	10.13 ± 2.36	19.40 ± 7.81	<0.001	0.610 (0.497–0.740)
CLI	12.34 ± 2.06	11.65 ± 2.04	10.81 ± 2.08	15.16 ± 2.19	<0.001	0.610 (0.505–0.728)
FKGL	16.84 ± 8.87	10.40 ± 2.29	9.77 ± 1.85	16.40 ± 6.26	<0.001	0.492 (0.348–0.671)
GFI	19.69 ± 9.07	13.16 ± 2.58	12.77 ± 2.13	18.85 ± 6.42	<0.001	0.453 (0.280–0.654)
SMOG	15.30 ± 3.75	12.42 ± 1.84	12.11 ± 1.58	15.59 ± 3.25	<0.001	0.443 (0.268–0.640)
FRES	35.68 ± 23.96	51.38 ± 12.09	56.70 ± 10.28	31.24 ± 19.00	<0.001	0.502 (0.351–0.689)

Values are presented as mean ± SD. Higher ARI, CLI, FKGL, GFI, and SMOG scores indicate greater estimated reading difficulty, whereas higher FRES scores indicate easier readability.

Discussion

Principal Findings

In this multidimensional comparative evaluation of four LLM chatbots, most responses to caregiver-facing pediatric burn questions were classified as safe and received generally favorable accuracy ratings, although differences were observed across several evaluated domains. ChatGPT had the highest observed proportion of safe responses, whereas Doubao had the lowest. Median accuracy was high across all models, and the statistically significant difference in accuracy appeared to reflect variation in completeness and precision rather than substantive disagreement with guideline-based recommendations. Empathy differed more markedly, with Doubao scoring lower than the other models. ChatGPT and Gemini generally achieved higher DISCERN and EQIP scores, whereas JAMA scores were uniformly low and did not differ across models. Readability was suboptimal overall: Gemini and DeepSeek were comparatively more readable, but grade-level indices remained above the levels typically recommended for patient-facing health materials [29-31].

These findings underscore why LLMs used in high-risk caregiver contexts should not be judged by apparent fluency alone. In pediatric burn care, a response may be medically plausible but still insufficient if it does not identify red flags, fails to discourage non-recommended topical agents or home remedies, obscures urgent action steps within lengthy explanatory text, or provides reassurance without appropriate escalation advice [5-8]. The present study therefore provides a multidimensional evaluation that integrates safety, accuracy, empathy, information quality, and readability—dimensions that collectively determine whether chatbot advice is usable by families during acute and inpatient burn care [18,19].

Safety and Accuracy in Pediatric Burn First Aid

Although most responses were classified as safe, the remaining potentially unsafe outputs warrant attention. Because pediatric burn injuries require time-sensitive care, even an occasional omission of urgent referral criteria or endorsement of inappropriate practices may delay assessment or worsen the injury. Safety-critical response elements should include stopping the burning process, cooling the burn wound with cool running water, removing clothing and jewelry when safe while leaving adherent material in place, loosely covering the burn wound with clean, non-adherent material, avoiding ice and non-recommended topical agents or home remedies, and stating clear indications for urgent medical assessment [5-10].

Although median accuracy was similar across models, the small but significant model effect suggests that differences were driven by completeness, wording precision, and the prominence of red-flag information rather than by substantive disagreement with guidelines. For caregivers, accuracy requires not only correct facts but also recognition of high-risk features, including involvement of the face, neck, hands, feet, genitalia, or perineum; a large burn size expressed as a percentage of total body surface area; suspected full-thickness burn injury; infancy; a chemical or electrical mechanism; suspected inhalation injury; or possible non-accidental injury [5-8]. Future scoring systems should therefore weight omissions of red flags and unsafe sequencing more heavily than minor stylistic differences.

Empathy and Risk Communication

Empathy showed one of the largest inter-model differences. This domain is particularly relevant for pediatric burns because caregivers may experience panic, guilt, or uncertainty immediately after injury. Effective communication should briefly acknowledge distress, avoid blame, and then transition to clear actions. Excessive reassurance may be problematic if it reduces the perceived urgency, whereas an overly directive or impersonal style may reduce trust or adherence. The ideal chatbot response should combine brief emotional acknowledgment with actionable triage and explicit guidance on when urgent medical assessment is required [13].

These results suggest that strategies for generating empathic responses in medical LLMs should be integrated with risk-communication templates rather than added as generic introductory phrases. In pediatric burn scenarios, empathic language should support caregivers in carrying out burn first aid, recognizing warning signs, and obtaining timely medical assessment; it should not substitute for clinical risk stratification.

Information Quality and Transparency

ChatGPT and Gemini achieved higher DISCERN and EQIP scores, suggesting stronger performance in structure, completeness, and usability of health-education content [20-22]. Nevertheless, JAMA scores were low and similar across all models. This pattern reflects an important limitation of general-purpose

chatbots: they may generate coherent and useful text while failing to disclose authorship, cite sources, specify currency, or provide traceable links to supporting evidence [23].

Source transparency is particularly important in caregiver-facing burn guidance. Families and clinicians need to distinguish guideline-derived advice from general, model-generated content. Future applications should incorporate retrieval-augmented generation based on authoritative burn care guidelines, traceable source links, last-updated dates, and localized escalation prompts. Responses should also clearly state that the information is intended for general education and does not replace professional assessment, especially when injury severity, anatomical site, age, mechanism, or symptoms indicate high risk [16-19].

Readability and Caregiver Usability

Readability was one of the most practically important findings. During the first minutes after a pediatric burn, caregivers are likely to be under stress and may have limited capacity to process long, technical, or densely qualified explanations. Even accurate advice may fail if the user cannot quickly identify what to do first. ChatGPT responses had comparatively high estimated reading-grade levels, indicating potential barriers to rapid comprehension. Doubao had relatively high reading-grade estimates and lower empathy ratings, suggesting limitations across both linguistic accessibility and supportive communication.

LLM outputs for pediatric burn care should adopt a layered, action-first format. The first layer should provide a concise emergency checklist: stop the burning process, cool the burn wound with cool running water, remove clothing and jewelry when safe while leaving adherent material in place, loosely cover the burn wound with clean, non-adherent material, avoid ice and non-recommended topical agents or home remedies, and obtain urgent medical assessment when specified red flags are present. The second layer can explain the rationale for these steps, and the third layer can provide guidance on inpatient burn care, follow-up care, and rehabilitation. Parent-facing outputs should use short sentences, numbered steps, prominent warnings, and low reading-grade levels by default [29-31]. Future evaluations should supplement formula-based readability metrics with caregiver comprehension studies and task-performance testing.

The interplay between readability, empathy, and safety in LLM-generated burn care responses deserves particular attention. Our findings suggest that models producing more empathetic language (e.g., Gemini, which received the highest empathy scores) did not consistently achieve the most favorable readability scores, while models with lower empathy ratings (e.g., Doubao) also produced some of the most difficult-to-read text. This pattern highlights a fundamental challenge in AI-assisted health communication: achieving a response that is simultaneously concise, action-oriented, supportive, and comprehensible to a distressed caregiver with limited health literacy remains elusive. Future model development and prompt engineering for pediatric burn guidance should prioritize this intersection—designing responses that lead with clear, prioritized action steps; use plain, concrete language; and embed empathetic acknowledgment without diluting or obscuring critical safety information.

Implications for Clinical Practice and Model Development

The findings of this study do not support the use of general-purpose chatbots as a standalone source of clinical guidance during pediatric burn emergencies. However, LLMs may become useful adjuncts if embedded within clinically governed systems that enforce safety-critical content, identify high-risk prompts, present advice in health-literacy-appropriate formats, and direct users to professional assessment when clinical uncertainty remains [16-19].

For developers, pediatric burn prompts may be designed to activate condition-specific safety templates rather than generic health-advice responses.

Strengths and Limitations

This study has several strengths. The evaluation was grounded in pediatric burn guidelines and patient-education standards; it compared multiple widely accessible chatbots; it assessed several complementary domains rather than focusing on factual accuracy alone; and it used blinded expert raters with high inter-rater agreement. The use of multiple readability indices also enabled a broader assessment of textual accessibility.

Several limitations should be acknowledged. First, only four general-purpose chatbots were evaluated, and their performance may change as models and interfaces are updated. Second, each question was submitted only once to each model, and repeated querying may have produced different responses. Third, all questions were submitted in English, although caregivers may seek information in other languages or dialects. Fourth, the study used hypothetical caregiver-facing questions rather than real emergency consultations, and comprehension or task performance among actual caregivers was not assessed. Fifth, formula-based readability indices do not fully capture layout, numeracy, emotional state, health literacy, or the ability to act on information under stress. Sixth, although evaluators were blinded to model identity and inter-rater agreement was high, expert ratings remain partly subjective. Finally, the findings represent model performance during a specific query period and under specified access conditions and should not be interpreted as stable rankings.

Conclusions

In this comparative evaluation, most responses generated by four LLM chatbots to caregiver questions about pediatric burns were classified as safe and received generally favorable accuracy ratings. Nevertheless, potentially unsafe content, variable empathy, limited source transparency, and substantial readability barriers remained. LLM-generated information should not be used as the sole source of guidance during pediatric burn emergencies. Effective caregiver-facing systems should provide concise and prioritized burn first aid instructions, clearly identify warning signs requiring urgent assessment, use plain language, and distinguish general education from individualized clinical advice. Clinical oversight, authoritative evidence retrieval, and condition-specific safety safeguards would be important for the responsible use of LLMs in pediatric burn education and caregiver support. Specifically, we recommend that future pediatric burn information systems incorporate three core design features: (1) an action-first emergency checklist visible at the top of any burn-related response; (2) mandatory red-flag statements

for high-risk scenarios; and (3) a plain-language summary of key safety warnings, with clear links to authoritative sources and explicit guidance on when to seek urgent professional care.

Abbreviations

AI, artificial intelligence; CHART, Chatbot Assessment Reporting Tool; LLM, large language model.

Declarations

Ethics approval and consent to participate

Not applicable. In this study, raters only performed quality assessment on the answers. In accordance with the regulations on scientific research ethics in China[41], the Ethics Committee of Beijing Children's Hospital, Capital Medical University granted an exemption from ethical review. We have fully informed the raters of the content of this study and obtained written informed consent.

Consent for publication

Not applicable. This manuscript does not contain any identifiable individual person's data, images, or videos.

Availability of data and materials

All data generated or analysed during this study are included in this article and its supplementary information files. Specifically, the question-screening record, reference-answer mapping table, cleaned question-and-answer dataset, pairwise comparison results, qualitative safety analysis, and raw scoring data are provided in Supplementary Materials 1–6, respectively. The statistical code and additional analytical materials are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no financial or non-financial competing interests. None of the authors received research funding, consulting fees, honoraria, stock options, or other financial support from the developers or providers of ChatGPT, Gemini, DeepSeek, or Doubao in connection with this study.

Funding

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

CW contributed to the conception and design of the study, development of the caregiver-facing question set, data collection, data curation, interpretation of the results, and drafting of the original manuscript. YL

contributed to the study design, construction of the reference-answer mapping table, coordination of the expert evaluation, data validation, and critical revision of the manuscript. YM contributed to data organization, statistical analysis, preparation of figures and tables, and interpretation of the quantitative results. ZW supervised the study, provided clinical and methodological guidance, resolved disagreements during question screening and reference-standard development, and critically revised the manuscript for important intellectual content. QY contributed to the development and validation of the evaluation framework, expert scoring, quality control, and revision of the manuscript. All authors read and approved the final manuscript and agreed to be accountable for all aspects of the work.

Acknowledgements

The authors thank Yanni Wang, Kun Chen, and Yanhua Liu for their contributions to the evaluation process. The authors also thank Jigang Chen, MD, from the Department of Burn Surgery for providing clinical validation of the reference-answer mapping table and safety classification criteria.

References

1. World Health Organization. Burns [Internet]. Geneva, World Health Organization. : 2023 Oct 13 [cited 2026 Jul 5]. Available from: <https://www.who.int/news-room/fact-sheets/detail/burns>
2. Peck MD. Epidemiology of burns throughout the world. Part I: distribution and risk factors. *Burns*. 2011;37(7):1087–100. 10.1016/j.burns.2011.06.005.
3. Brusselaers N, Monstrey S, Vogelaers D, Hoste E, Blot S. Severe burn injury in Europe: a systematic review of the incidence, etiology, morbidity, and mortality. *Crit Care*. 2010;14(5):R188. 10.1186/cc9300.
4. Krishnamoorthy V, Ramaiah R, Bhananker SM. Pediatric burn injuries. *Int J Crit Illn Inj Sci*. 2012;2(3):128–34. 10.4103/2229-5151.100889.
5. ISBI Practice Guidelines Committee. ISBI practice guidelines for burn care. *Burns*. 2016;42(5):953–1021. 10.1016/j.burns.2016.05.013.
6. British Burn Association. First aid clinical practice guidelines [Internet]. 2018 Apr [cited 2026 Jul 5]. Available from: <https://www.britishburnassociation.org/wp-content/uploads/2017/06/BBA-First-Aid-Guideline-24.7.18.pdf>
7. American Burn Association. Guidelines for burn patient referral [Internet]. Chicago: American Burn Association. 2022 [cited 2026 Jul 5]. Available from: <https://www.ameriburn.org/burn-care-team/resources/guidelines-for-burn-patient-referral>
8. The Royal Children's Hospital Melbourne. Burns - acute management [Internet]. Melbourne: The Royal Children's Hospital Melbourne; 2026 [updated 2026; cited 2026 Jul 5]. Available from: https://www.rch.org.au/clinicalguide/guideline_index/burns/
9. Griffin B, Cabilan CJ, Ayoub B, Xu HG, Palmieri T, Kimble R, et al. The effect of 20 minutes of cool running water first aid within three hours of thermal burn injury on patient outcomes: a systematic review and meta-analysis. *Australas Emerg Care*. 2022;25(4):367–76. 10.1016/j.auec.2022.05.004.

10. Wood FM, Phillips M, Jovic T, Cassidy JT, Cameron P, Edgar DW, et al. Water first aid is beneficial in humans post-burn: evidence from a bi-national cohort study. *PLoS ONE*. 2016;11(1):e0147259. 10.1371/journal.pone.0147259.
11. Atiyeh BS, Janom HH. Physical rehabilitation of pediatric burns. *Ann Burns Fire Disasters*. 2014;27(1):37–43.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–40. 10.1038/s41591-023-02448-8.
13. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–96. 10.1001/jamainternmed.2023.1838.
14. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthc (Basel)*. 2023;11(6):887. 10.3390/healthcare11060887.
15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. 10.1371/journal.pdig.0000198.
16. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613–22. 10.1038/s41591-024-03097-1.
17. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*. 2023;6(1):120. 10.1038/s41746-023-00873-0.
18. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open*. 2025;8(2):e2457879. 10.1001/jamanetworkopen.2024.57879.
19. The CHART. Collaborative. Reporting guideline for chatbot health advice studies: the CHART statement. *JAMA Netw Open*. 2025;8(8):e2530220. 10.1001/jamanetworkopen.2025.30220.
20. 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(2):105–11. 10.1136/jech.53.2.105.
21. Charnock D. The DISCERN handbook: quality criteria for consumer health information on treatment choices. Abingdon: Radcliffe Medical; 1998.
22. Moulton 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):165–75. 10.1111/j.1369-7625.2004.00273.x.
23. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: caveat lector et viewer—let the reader and viewer beware. *JAMA*. 1997;277(15):1244–5. 10.1001/jama.1997.03540390074039.

24. 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(9):2070–7. 10.1111/j.1572-0241.2007.01325.x.
25. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32(3):221–33. 10.1037/h0057532.
26. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Millington (TN): Chief of Naval Technical Training; 1975. Research Branch Report No. 8–75. 10.21236/ADA006655
27. Coleman M, Liao TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;60(2):283–4. 10.1037/h0076540.
28. McLaughlin GH. SMOG grading: a new readability formula. *J Read*. 1969;12(8):639–46.
29. Gunning R. The technique of clear writing. New York: McGraw-Hill; 1952.
30. Weiss BD. Health literacy and patient safety: help patients understand. Manual for clinicians. 2nd ed. Chicago (IL): American Medical Association Foundation; 2007.
31. Safeer RS, Keenan J. Health literacy: the gap between physicians and patients. *Am Fam Physician*. 2005;72(3):463–8.
32. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378–82. 10.1037/h0031619.
33. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–8. 10.1037/0033-2909.86.2.420.
34. Cochran WG. The comparison of percentages in matched samples. *Biometrika*. 1950;37(3–4):256–66. 10.1093/biomet/37.3-4.256.
35. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12(2):153–7. 10.1007/BF02295996.
36. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc*. 1937;32(200):675–701. 10.1080/01621459.1937.10503522.
37. Kendall MG, Babington Smith B. The problem of m rankings. *Ann Math Stat*. 1939;10(3):275–87. 10.1214/aoms/1177732186.
38. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1(6):80–3. 10.2307/3001968.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x.
40. R Core Team. R: a language and environment for statistical computing [Internet]. Vienna: R Foundation for Statistical Computing. 2025 [cited 2026 Jul 5]. Available from: <https://www.R-project.org/>

41. National Health and Family Planning Commission of the People’s Republic of China. Measures for the Ethical Review of Biomedical Research Involving Humans: Order No. 11 of the National Health and Family Planning Commission [EB/OL]. 2016 Oct 12 [cited 2026 Sep 9].
<https://www.nhc.gov.cn/wjw/c100221/202201/558653bd3d1e4a078904152a30605a03.shtml>

Figures

Figure 1. Study workflow for the comparative evaluation of large language model chatbots in pediatric burn first aid and inpatient care

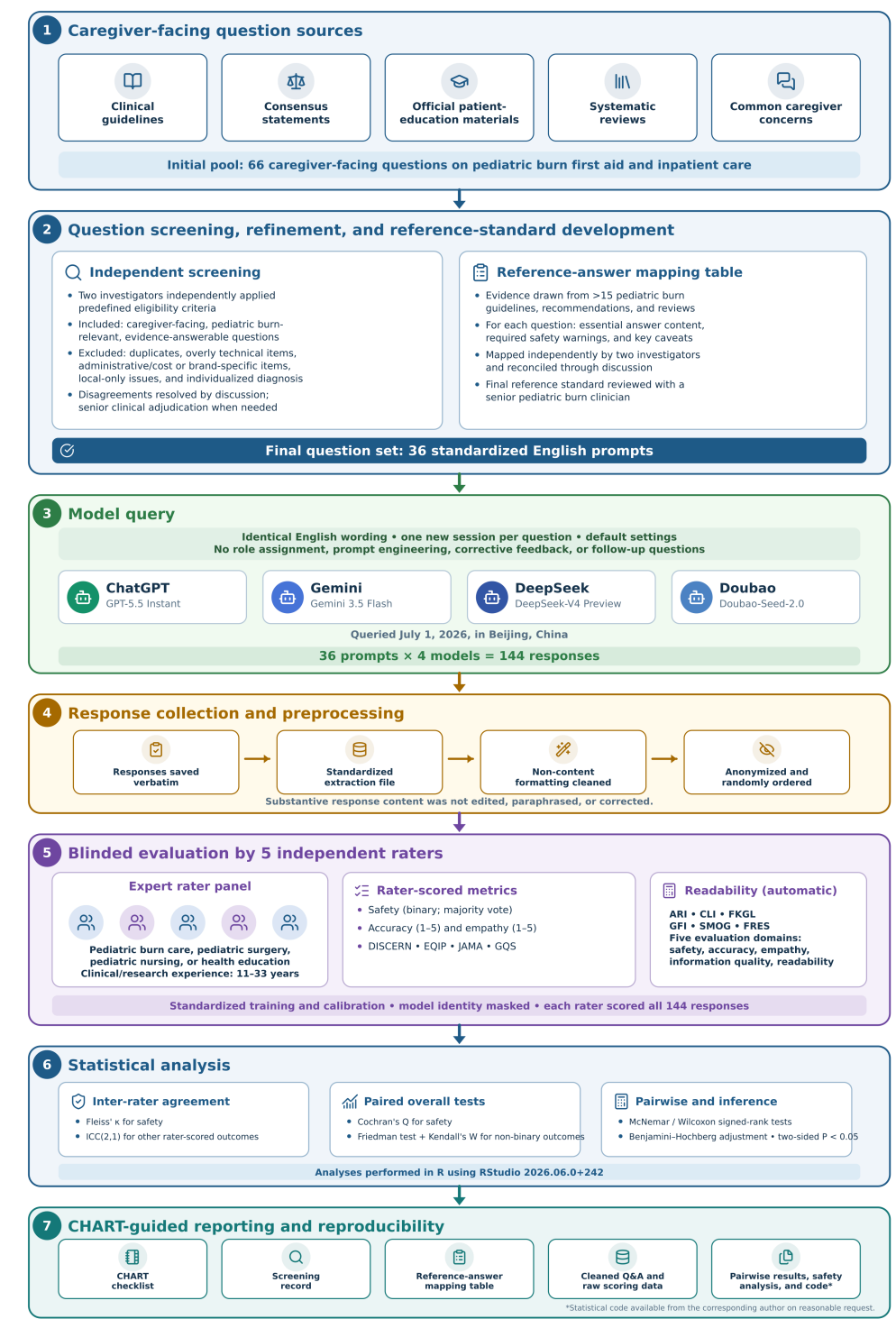


Figure 1

Study workflow for the cross-sectional comparative evaluation of the safety, accuracy, empathy, information quality, and readability of large language models in answering caregiver questions about pediatric burn first aid and inpatient care.

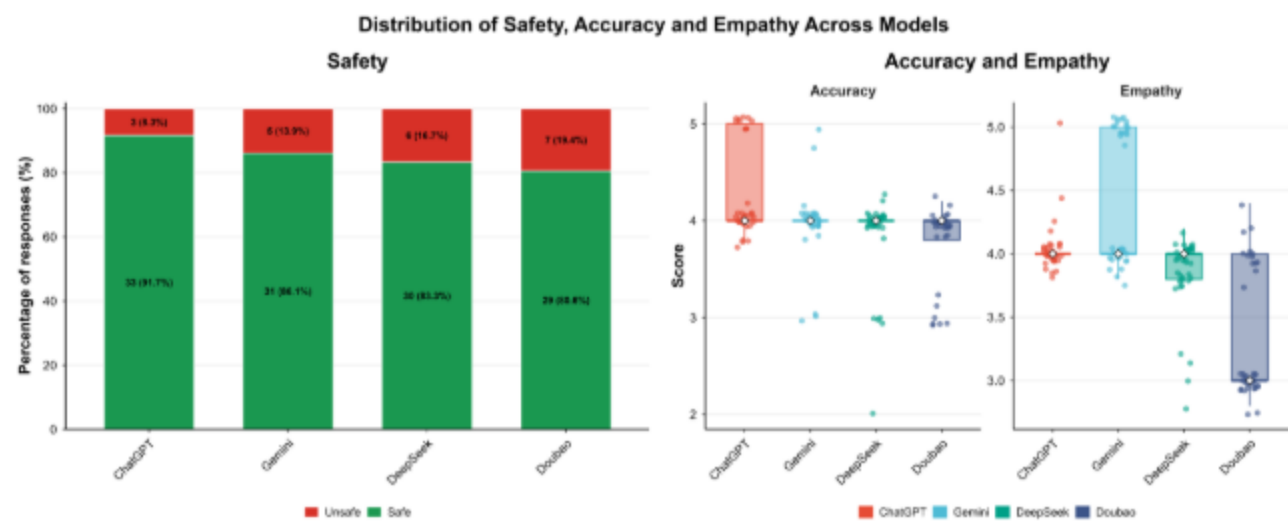


Figure 2

Distribution of safety, accuracy, and empathy across models.

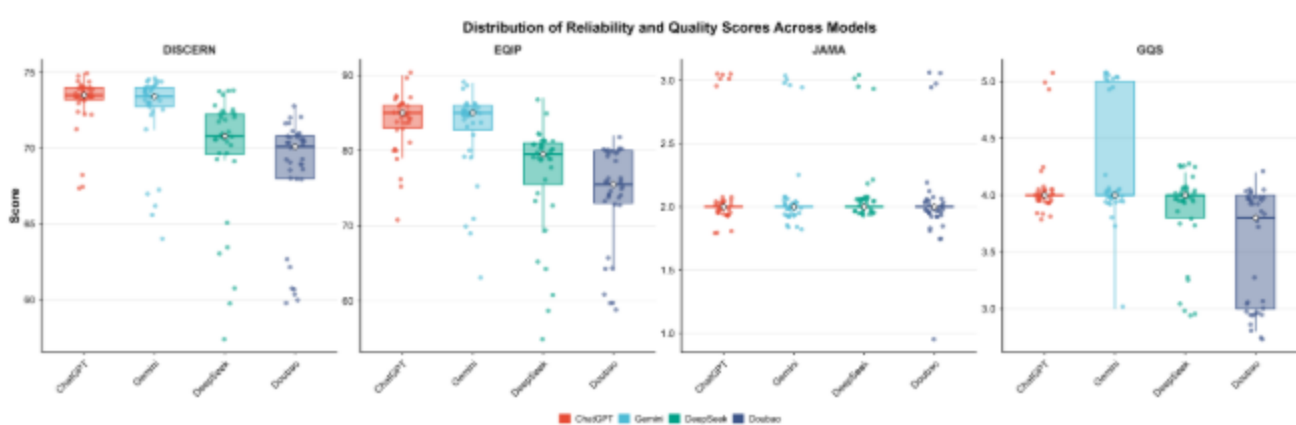


Figure 3

Distributions of DISCERN, EQIP, JAMA benchmark, and Global Quality Score across models.

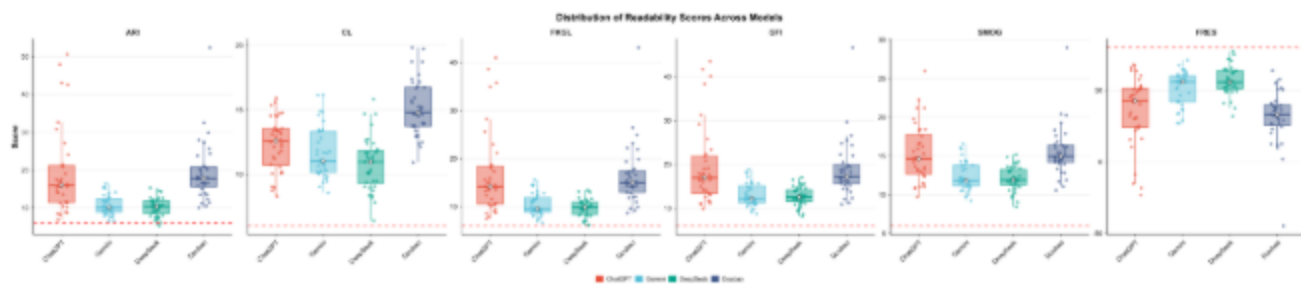


Figure 4

Distributions of readability indices across models.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial1.docx](#)
- [SupplementaryMaterial2.docx](#)
- [SupplementaryMaterial3.docx](#)
- [SupplementaryMaterial4.docx](#)
- [SupplementaryMaterial5.docx](#)
- [SupplementaryMaterial6.xlsx](#)