Original Paper

Real-World Use and Expert-Guided Optimization of a Large Language Model–Based Pediatric Health Assistant: Retrospective Observational Study

Qiuhong Wei[1,2,3,4,5], MD, PhD; Li Xiao[1,6], MD, PhD; Xin Zhang[1], PhD; Hong Wei[7], MD, PhD; Ruijue Wang[1,8], MS, MD; Suqiao Ou[1], MS, MD; Chao Zhang[9], MS; Huarong Huo[9], MS; Yu Peng[9], MS; Lishen Liu[9], BS; Ping Song[1*], BS; Ziyu Hua[7*], MD, PhD; Ximing Xu[1*], PhD

[1]Big Data Center for Children's Medical Care, Children's Hospital of Chongqing Medical University, Chongqing, China

[2]Department of Children's Healthcare, Children's Hospital of Chongqing Medical University, Chongqing, China

[3]National Clinical Research Center for Children and Adolescents' Health and Diseases, Chongqing, China

[4]Intelligent Application of Big Data in Pediatrics Engineering Research Center of Chongqing Education Commission, Chongqing, China

[5]Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China

[6]Department of Hematology, Children's Hospital of Chongqing Medical University, Chongqing, China

[7]Department of Neonatology, Children's Hospital of Chongqing Medical University, Chongqing, China

[8]Department of Hepatobiliary Surgery, Children's Hospital of Chongqing Medical University, Chongqing, China

[9]Zuoyi Tech, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Ximing Xu, PhD
Big Data Center for Children's Medical Care, Children's Hospital of Chongqing Medical University
No. 20 Jinyu Avenue, Liangjiang New Area
Chongqing, 401122
China
Phone: 86 18526547007
Email: ximing@hospital.cqmu.edu.cn

Abstract

Background: Pediatric care resources are limited and unevenly distributed in many health systems. Caregivers increasingly turn to large language models (LLMs) for children's health information, but real-world evidence on LLM use and response quality in general pediatric care is limited.

Objective: This study aimed to characterize real-world use, consultation needs, and response quality of Xiaoyi, an LLM-based pediatric health assistant, and to evaluate the impact of expert-guided optimization on response quality.

Methods: This study collected all consultation sessions submitted between April 7 and December 31, 2025, to Xiaoyi, a pediatric health assistant co-developed by Children's Hospital of Chongqing Medical University and Zuoyi Tech. Temporal patterns of consultation activity were characterized; all sessions were annotated for pediatric subspecialty, reason for consultation, and presenting symptoms through LLM-assisted labeling with pediatrician validation. A multidisciplinary panel of 40 pediatricians evaluated Xiaoyi's response quality over 2 rounds using predefined criteria established by expert consensus: an initial evaluation informed system optimization (prompt revision, knowledge-base and tool expansion, and validation rules) and a follow-up evaluation assessed its impact. In each round, 1000 sessions randomly sampled from the corresponding pre- or postoptimization period (2000 in total) were rated at the session level (overall quality and clinical appropriateness) and the turn level (6 predefined deficiency categories). Categorical ratings were compared using the chi-square test, and turn-level deficiency rates were compared using generalized estimating equations to account for within-session correlation.

Results: Caregivers of 8266 children (median age 4.2, IQR 1.3-8 years; 4705/8266, 56.9% male) from 31 of 34 provincial-level divisions in China initiated 16,074 sessions, of which 7046 (43.8%) occurred outside routine clinic hours (08:00-17:30). Queries spanned 31 subspecialties, most frequently child health care (2025/16,074, 12.6%), pulmonology (1903/16,074, 11.8%), and gastroenterology (1833/16,074, 11.4%). The leading reasons for consultation were symptom consultation (7845/16,074, 48.8%), medical report interpretation (2531/16,074, 15.7%), and triage guidance (2515/16,074, 15.6%); among symptom consultation

XSL•FO
RenderX

sessions, fever (2719/16,074, 34.7%) and cough (2349/16,074, 29.9%) were most common. After expert-guided optimization, the number of sessions rated good (n=1000) in overall quality (the response fully resolving the caregiver's question with actionable information) increased from 652 (65.2%) to 743 (74.3%), and the number rated clinically appropriate increased from 614 (61.4%) to 874 (87.4%, both $P<.001$). The turn-level deficiency rate declined from 19.4% (790/4073 turns) to 6.9% (236/3419 turns; odds ratio 0.30, 95% CI 0.24-0.37; $P<.001$), with suboptimal clinical recommendations (not optimally tailored to the child's age or clinical context) remaining the most frequent category.

Conclusions: This real-world study suggests that an LLM-based health consultation assistant may complement existing pediatric care, with expert-guided optimization substantially improving response quality. However, the persistence of deficiencies highlights the need for continued monitoring and refinement to ensure clinical reliability.

(J Med Internet Res 2026;28:e105346) doi: 10.2196/105346

KEYWORDS

large language model; pediatrics; health consultation; real-world study; expert-guided optimization

Introduction

In pediatric practice, young children often cannot articulate their symptoms, and the initial recognition of illness and the decision to seek care rely primarily on caregivers. During this process, caregivers frequently experience anxiety and may misinterpret clinical warning signs [1]. Moreover, pediatric care resources are limited in many health systems, with shortages of pediatricians and uneven geographic distribution of services, particularly in low- and middle-income settings [2]. Caregiver concerns also frequently arise outside routine clinic hours, and existing online consultation services are often hindered by delayed responses and variable quality [3,4]. Common caregiver concerns include respiratory, allergic, and skin symptoms, which may be associated with environmental exposures such as air pollution, tobacco smoke, aeroallergens, and indoor mold [5]. Providing caregivers with accessible, low-cost, and reliable pediatric health consultations beyond traditional clinical settings has therefore emerged as a pediatric health priority.

Recent advances in large language models (LLMs) offer new possibilities for addressing these challenges [6]. Owing to their capacity for natural language understanding, integration of multisource medical knowledge, and contextual reasoning, LLMs have been increasingly explored in real-world clinical care. In outpatient settings, LLM-based systems have improved consultation efficiency and facilitated referral from primary to specialty care [7,8]. In patient education, LLM-assisted chatbots and LLM-generated clinical documents have enhanced accuracy, economic efficiency, and patient comprehension [9,10]. LLMs have also shown promise in mental health interventions [11,12]. However, these applications have focused largely on adult populations, with limited attention to pediatric care.

Current evidence for LLM applications in pediatrics derives largely from case vignettes [13], examination questions [14], or simulated clinical scenarios [15], with few studies examining real-world deployment [16]. The available real-world studies have each addressed a single condition or population: AI-generated educational materials for developmental dysplasia of the hip [17], and a ChatGPT-delivered physical activity intervention for children with autism spectrum disorder [18]. Although these studies provide preliminary support for pediatric LLM applications, they are limited by small samples, controlled settings, and a narrow focus on specific diseases or populations. Consequently, the real-world use patterns and response quality of LLM-based systems in general pediatric care remain inadequately understood.

To address this gap, we analyzed real-world use data from Xiaoyi (小乙医生), a publicly deployed pediatric health assistant. This study aimed to: (1) characterize the temporal patterns and user distribution of pediatric LLM consultations; (2) delineate the landscape of caregivers' consultation needs, including subspecialty distribution, reasons for consultation, and symptom spectrum; and (3) evaluate the quality of LLM-generated responses and the impact of expert-guided iterative optimization on response quality.

Methods

Study Design and Data Source

This was a retrospective observational study based on real-world use data from Xiaoyi (小乙医生), an LLM-based pediatric health assistant developed by Children's Hospital of Chongqing Medical University and Zuoyi Tech. Xiaoyi was deployed as a mini program within WeChat (Tencent), the most widely used social platform in China, with over 1 billion monthly active users. The platform is publicly accessible free of charge, with no enrollment, screening, or eligibility assessment; all consultations were user-initiated. As a routine public service, the platform was promoted through posts on the hospital's official WeChat account and posters within the hospital at launch (April 7-9, 2025), during National Nutrition Week (May 16, 2025), and around International Children's Day (June 4, 2025); these promotions targeted the general public and involved no individual invitation to participate in research. We analyzed all consultation data submitted between April 7, 2025 (the platform's launch date) and December 31, 2025.

Prior to use, caregivers were required to register a child health profile, with a single user permitted to register multiple children. Each child was assigned a unique profile ID, and each user ID was tied to the caregiver's WeChat account rather than to a device, so access from multiple devices did not create duplicate records. Registration collected only a nickname, the child's sex, and date of birth, with no identifiable information such as legal name, national identification number, or contact details. At registration, caregivers accepted a disclaimer stating that all personal data would be stored in deidentified form and used

XSL•FO
RenderX

solely for service provision and research, and that AI-generated responses were intended for health guidance only and did not constitute a medical diagnosis or substitute for professional consultation. Geographic origin was inferred from the registration IP address and used only at the provincial level; no overseas IP addresses were recorded. For each consultation, the platform logged a deidentified user ID, child profile ID, timestamp, caregiver text input, and corresponding AI response. Before analysis, all free-text content was screened for directly identifying information (eg, personal names, telephone numbers, identification numbers, and addresses), which was removed; all subsequent processing, including LLM-assisted annotation and expert evaluation, used this deidentified text.

Ethical Considerations

The study was approved by the Ethics Committee of Children's Hospital of Chongqing Medical University (approval number 2026-009), which reviewed both the collection of platform data and its retrospective use for research. Given the retrospective design and the exclusive use of deidentified data, the requirement for individual informed consent was waived. All data were stored on hospital servers, with access restricted to approved members of the research team. This study is reported in accordance with the Chatbot Assessment Reporting Tool (CHART) statement for chatbot health advice studies (Table S1 in Multimedia Appendix 1) [19].

The Xiaoyi System

Xiaoyi (小乙医生) is an LLM-based pediatric health assistant that processes each consultation through 3 sequential stages: multimodal input processing, clinical reasoning with tool invocation, and response delivery with auxiliary outputs (Figure 1). Instructions for accessing the system and a brief video overview are provided in Method S1 in Multimedia Appendix 1. Its official Chinese product name, "儿科AI家庭医生" (Pediatric AI Family Doctor), denotes an accessible, home-based source of health guidance for families; Xiaoyi provides health guidance only and does not make diagnoses or replace clinical care.

Caregivers initiated consultations via text, speech, or image. Speech inputs were transcribed by an automatic speech recognition module, and images (eg, laboratory or imaging reports and photographs of skin lesions) were converted by a general-purpose vision-language model (Qwen3-VL-32B), not specifically trained on dermatological images, into natural-language descriptions of observable features, leaving all clinical reasoning to the reasoning core; transcribed and parsed content was concatenated with any caregiver-provided text into a unified input for downstream processing.

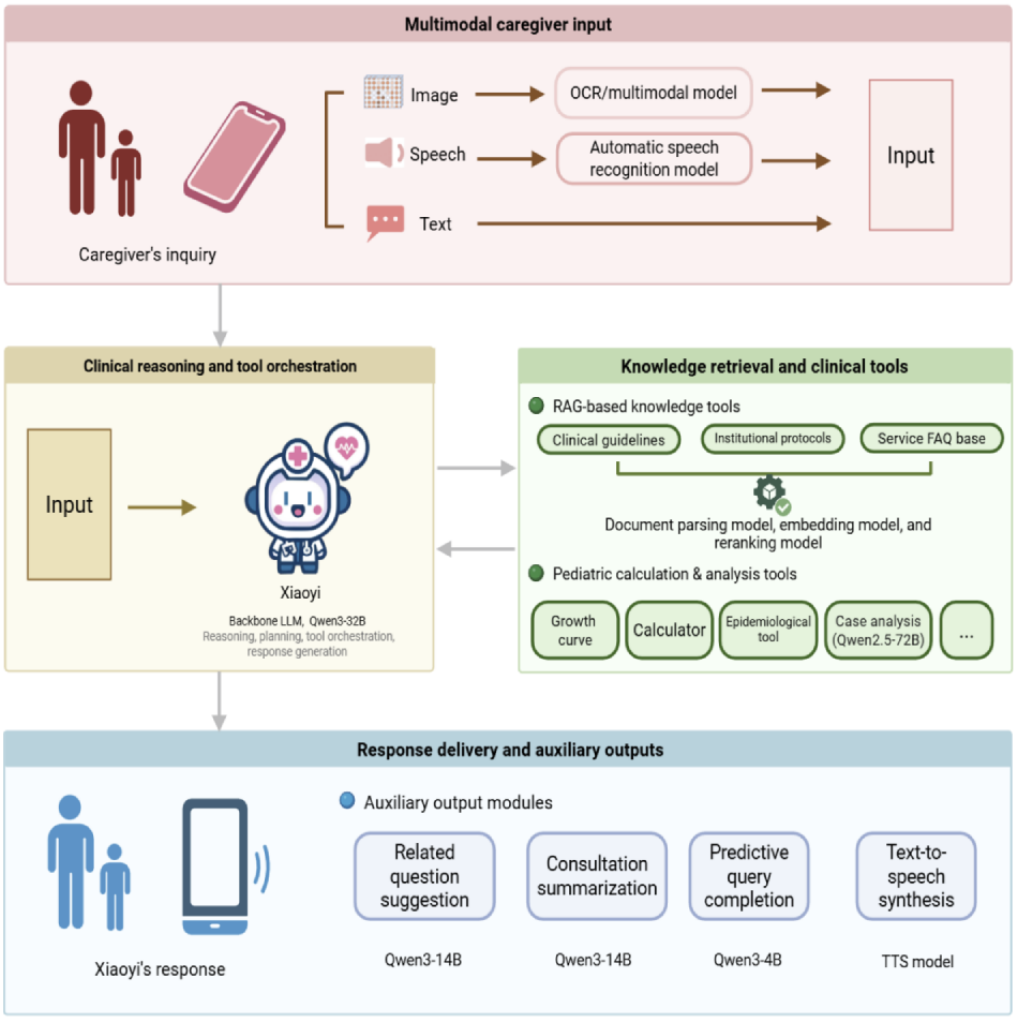
The primary reasoning core was a Qwen3-32B LLM during this study period. Each query was processed together with the child's platform profile (sex, date of birth, and prior consultations on the platform) as context; the system had no connection to the hospital electronic medical record. The model operated in a reasoning-and-acting (ReAct) loop, alternating between internal reasoning and external tool calls until sufficient evidence was gathered. Available resources comprised 3 retrieval-augmented knowledge bases and a suite of deterministic clinical tools. The knowledge bases covered clinical guidelines, departmental diagnostic and treatment protocols from our hospital, and a patient-service frequently asked questions repository, all sharing a unified pipeline of document parsing, embedding, and reranking. Examples of the clinical tools included a growth-curve module referenced to World Health Organization (WHO) standards, a calculator for pediatric indices such as BMI, a regional epidemiology tool that retrieved local infectious disease activity (without environmental data such as air quality indices or pollen counts) to support diagnostic reasoning during seasonal outbreaks, and a case-analysis tool powered by a Qwen2.5-72B model for complex diagnostic reasoning; the growth-curve module was introduced during optimization (Method S3 in Multimedia Appendix 1).

Four lightweight modules subsequently ran in parallel to support related-question suggestion, consultation summarization (both Qwen3-14B), predictive query completion (Qwen3-4B), and speech synthesis.

All deployed models were adapted to pediatric clinical contexts through instruction tuning, task-specific prompt engineering, and curated knowledge bases, rather than being used off-the-shelf. All models ran on local institutional infrastructure, so consultation content was not transmitted to external model providers during platform operation.

XSL•FO
RenderX

Figure 1. Architecture of the Xiaoyi Pediatric Health Consultation System. Caregiver inputs are processed through clinical reasoning and tool orchestration by a backbone large language model, which invokes knowledge retrieval and pediatric clinical tools to generate responses with auxiliary outputs. LLM: large language model; OCR: optical character recognition; RAG: retrieval-augmented generation; Service FAQ base: patient-service frequently asked questions repository; TTS: text-to-speech.



Consultation Characterization and Coding

Consultation data were organized into 2 levels. A turn was the basic unit of interaction, defined as a single question-answer pair comprising one caregiver input and the one corresponding response generated by Xiaoyi. A session was a distinct conversational thread on the platform, beginning when a caregiver initiated a new consultation and ending when the caregiver started a new thread or the conversation was closed, and encompassing all turns exchanged within that thread. All consultation sessions were systematically annotated along 3 dimensions: pediatric subspecialty, reason for consultation, and presenting symptoms.

Subspecialty and reason for consultation were assigned through LLM-assisted annotation with pediatrician validation. Specifically, DeepSeek-V3.2, accessed via the application programming interface (model identifier, deepseek-chat, in nonthinking mode) on February 9, 2026, analyzed the full text of each session and assigned it to the single most relevant category within two predefined taxonomies: (1) pediatric subspecialty, comprising 31 categories spanning medical, surgical, and auxiliary specialties (Table S2 in Multimedia Appendix 1); and (2) reason for consultation, comprising 5 categories (symptom consultation, medical report interpretation, triage guidance, medication counseling, and health information inquiry; Table S3 in Multimedia Appendix 1). Both tasks used structured prompts that embedded the complete classification taxonomies and decision rules (Methods S2 in Multimedia Appendix 1). To assess annotation accuracy, the full set of 16,074 sessions was ordered chronologically by session timestamp to form a sampling frame, from which 1000 sessions were drawn by simple random sampling without replacement for each task. Two pediatricians (WQH and XL) independently reviewed the 1000 sampled sessions, with disagreements adjudicated by a third senior pediatrician (HZY). The proportion of assigned symptom labels discordant with the final pediatrician consensus was 3% (95% CI 2%-4.2%) for subspecialty classification and 4% (95% CI 2.9%-5.4%) for consultation reason classification.

For sessions classified as symptom consultation, we built a symptom lexicon using a data-driven approach to accommodate the variable, colloquial wording used by caregivers. Caregiver inputs were processed by the same model (DeepSeek-V3.2), which identified symptom mentions and grouped them into

XSL•FO
RenderX

standard terms based on meaning (eg, “发烧”, “高烧”, and “烧不退” were all mapped to “发热” [fever]). A pediatrician (WQH) then reviewed the lexicon entry by entry, removing nonsymptom terms, correcting misclassifications, and organizing entries by subspecialty. The final lexicon contained 43 standard symptoms and 241 expression variants (Table S4 in Multimedia Appendix 1). Symptom mentions in session text were then matched to standard terms using regular expressions; negated mentions (eg, “no fever”) were excluded by the regular expressions. To assess matching accuracy, 1000 sessions were drawn from the 7845 symptom-consultation sessions by simple random sampling without replacement and reviewed by the same panel (WQH, XL, and HZY); the LLM-assigned labels were discordant with pediatrician consensus in 3.1% (95% CI, 2.1%-4.4%) of cases.

Expert Evaluation Framework

To support iterative quality improvement, we established an expert evaluation framework in which a multidisciplinary panel of pediatricians reviewed model outputs to identify recurring deficiencies and guide system optimization. Two rounds of evaluation were conducted during the study period. Based on findings from the initial evaluation, the research team refined the Xiaoyi system through prompt revisions, expansion of the retrieval knowledge base and tools, and introduction of validation rules (Method S3 in Multimedia Appendix 1). The updated system was then redeployed, and a follow-up evaluation was conducted to assess the impact of these refinements.

The evaluation panel comprised 40 pediatricians, each holding the rank of attending pediatrician or above with at least 5 years of clinical experience, collectively covering all 31 pediatric subspecialties at our hospital (eg, neonatology and pulmonology). Prior to evaluation, a calibration meeting was held to standardize assessment criteria, introduce the scoring rules, and establish consensus on each evaluation dimension through joint review of example cases.

The initial evaluation, performed in July 2025, sampled sessions submitted between April 7 and June 30, 2025; expert feedback informed system optimization, which was completed on July 31, 2025. The follow-up evaluation, performed in January 2026, sampled sessions submitted to the optimized system between August 1 and December 31, 2025. Sessions from July 2025 (the optimization period) were excluded from both evaluations to ensure a clean separation between pre- and postoptimization periods. For each evaluation round, 1000 sessions were drawn from the corresponding period by simple random sampling without replacement. The 2 samples were therefore independent and represented their respective periods rather than the full study period. Each session was assigned to and rated by a single pediatrician within the matching subspecialty, without duplicate rating or consensus scoring; when multiple pediatricians covered the same subspecialty, sessions were randomly allocated among them with approximately balanced workload. For sessions involving uncertainty or complex clinical presentations, evaluators could consult senior colleagues within their subspecialty before finalizing the assessment.

Evaluation was conducted at 2 levels. At the session level, 2 indicators assessed Xiaoyi’s overall performance: overall quality, rated as good, fair, or poor; and clinical appropriateness, rated as appropriate, mildly inappropriate, moderately inappropriate, or severely inappropriate. At the turn level, each question-answer pair was evaluated for the presence of six predefined deficiency categories: (1) failure to understand the user’s intent; (2) noncontributory follow-up queries; (3) omission of critical history-taking; (4) premature diagnostic closure; (5) inadequate diagnostic reasoning; and (6) suboptimal clinical recommendations. Each deficient turn was assigned to the single most applicable category, with deficiencies outside these categories classified as “other.” The 6 deficiency categories were developed through discussion between the pediatric expert panel and the engineering team, drawing on common error patterns observed during prelaunch testing and on consensus regarding what constitutes a clinically meaningful deficiency in pediatric consultation. Complete definitions, decision rules, and representative examples for each rating level and deficiency category are provided in Table S5, Table S6, and Table S7 in Multimedia Appendix 1.

Statistical Analysis

All analyses were performed in R (version 4.4.2; R Foundation for Statistical Computing). Continuous variables were presented as mean (SD) or median (IQR) depending on their distribution, and categorical variables as frequencies and percentages. Group comparisons used the Mann-Whitney *U* test for continuous variables and the chi-square test for categorical variables. Turn-level deficiency rates were compared using generalized estimating equations (GEE) with an AR(1) working correlation structure to account for within-session correlation among sequential turns, with results reported as odds ratios (ORs) and 95% CIs. All tests were 2-sided, with statistical significance set at $P<.05$.

Results

Study Population

The Xiaoyi (小乙医生) platform was launched on April 7, 2025. As of December 31, 2025, 7243 users had registered and created health profiles for 8289 individuals (8266 children aged 18 years and 23 adults). As this study focused on pediatric care, adults were excluded from subsequent analyses. Caregivers of the 8266 children initiated 16,074 consultation sessions comprising 54,771 dialogue turns, with a mean of 3.4 (SD 3) turns per session (range 1-49).

Among the 8266 children, the median (IQR) age was 4.2 (1.3-8) years, and 4705 (56.9%) were male. Children resided in 31 of China’s 34 provincial-level administrative divisions, with Chongqing accounting for the largest proportion (5422/8266, 65.6%). Detailed distributions of age, sex, and geographic origin are presented in Table 1.

XSL•FO
RenderX

Table 1. Demographic characteristics of the study population (N=8266).

Characteristics	Values, n (%)
Sex	
Male	4705 (56.9)
Female	3561 (43.1)
Age group (years)	
Infancy (0-1)	1653 (20)
Toddler (1-3)	1769 (21.4)
Preschool (3-6)	1770 (21.4)
School age (6-12)	2439 (29.5)
Adolescence (12-18)	635 (7.7)
Region	
Chongqing	5422 (65.6)
Sichuan	1461 (17.7)
Guizhou	267 (3.2)
Others	1116 (13.5)

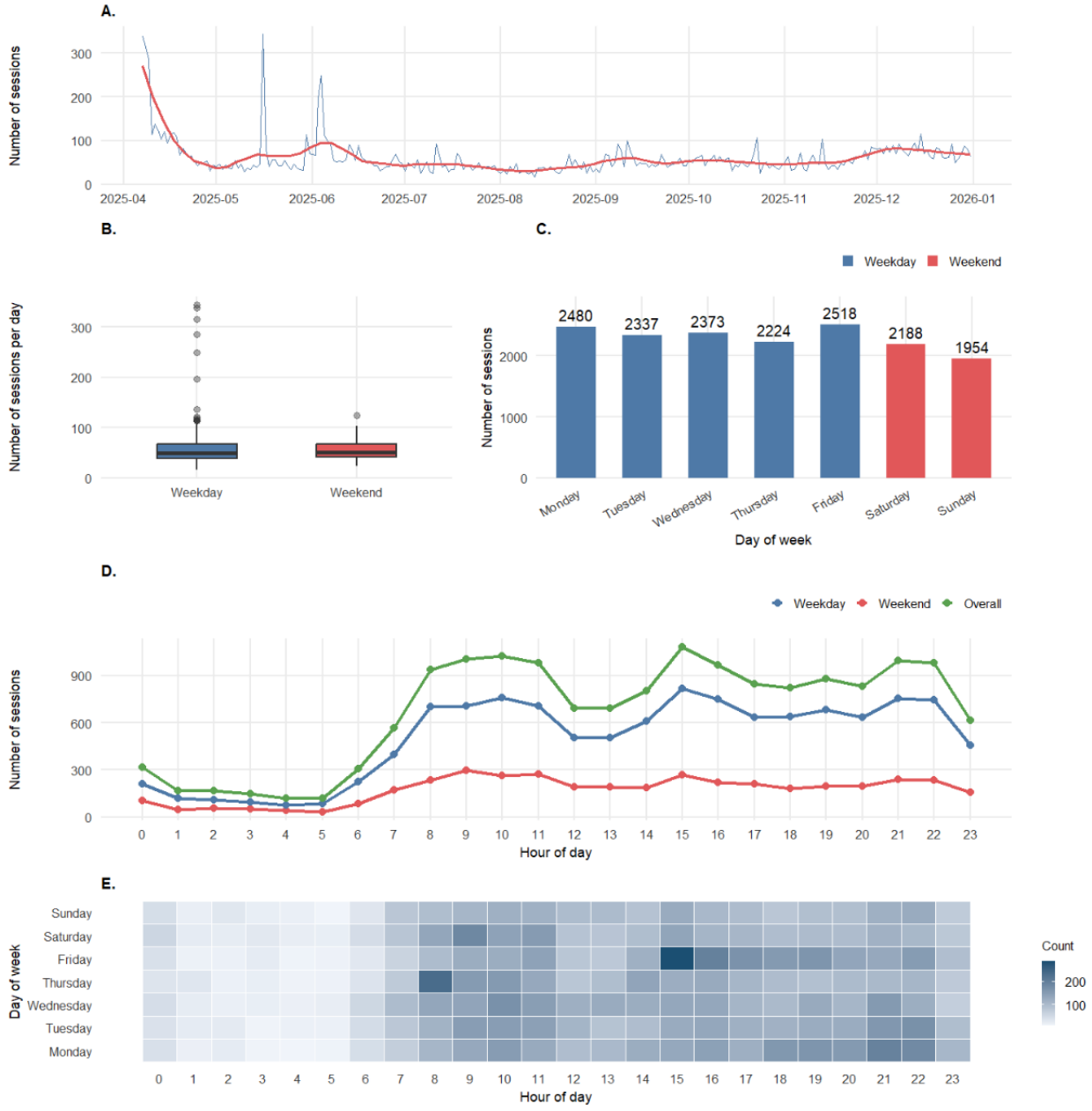
Temporal Patterns of Consultation Activity

Over the 269-day study period (April 7 to December 31, 2025), caregivers initiated 16,074 sessions, averaging 59.8 sessions per day (Figure 2A). Transient increases in daily volume (>200 sessions) occurred on April 7-9, May 16, and June 4, coinciding with the promotional activities described in the Methods. The median daily session count was similar between weekdays (49; IQR, 39-67) and weekends (51; IQR, 41-68; *P*=.94; Figures 2B and 2C).

Sessions were initiated across all 24 hours (Figures 2C and 2E). Daytime peaks occurred around 10:00 (n=1024) and around 15:00 (n=1082), followed by a third peak around 21:00 (n=994) within a sustained high-volume evening period (19:00-22:00). Overall, 7046 sessions (43.8%) occurred outside routine clinic hours (8:00-17:30), and hourly session counts between 0:00 and 6:00 ranged from 121 to 309.

XSL•FO
RenderX

Figure 2. Temporal patterns of consultation sessions during the study period. (A) Daily number of consultation sessions. (B) Daily consultation sessions on weekdays vs weekends. (C) Consultation sessions by day of week. (D) Consultation sessions by hour of day: weekdays, weekends, and overall. (E) Heatmap of consultation sessions by day of week and hour of day.



Clinical Characteristics of Consultation Sessions

Across the 16,074 sessions, caregiver queries spanned 31 pediatric subspecialties, with nonsurgical specialties accounting for 82.6% (13,102/15,872) and surgical specialties for 17.4% (2770/15,872). The 3 most frequently consulted subspecialties were child health care (2025; 12.6%), pulmonology (1903; 11.8%), and gastroenterology (1833; 11.4%). The age distribution of consulting children varied by subspecialty: child health care and gastroenterology consultations were predominantly for infants and toddlers, pulmonology consultations peaked during the toddler and preschool periods, and endocrinology consultations were concentrated in school-age children and adolescents. The full subspecialty ranking and the age distribution across the 10 most consulted subspecialties are shown in Figures 3A and 3B, and session counts for all subspecialties and consultation reasons are provided in Table S8 in Multimedia Appendix 1.

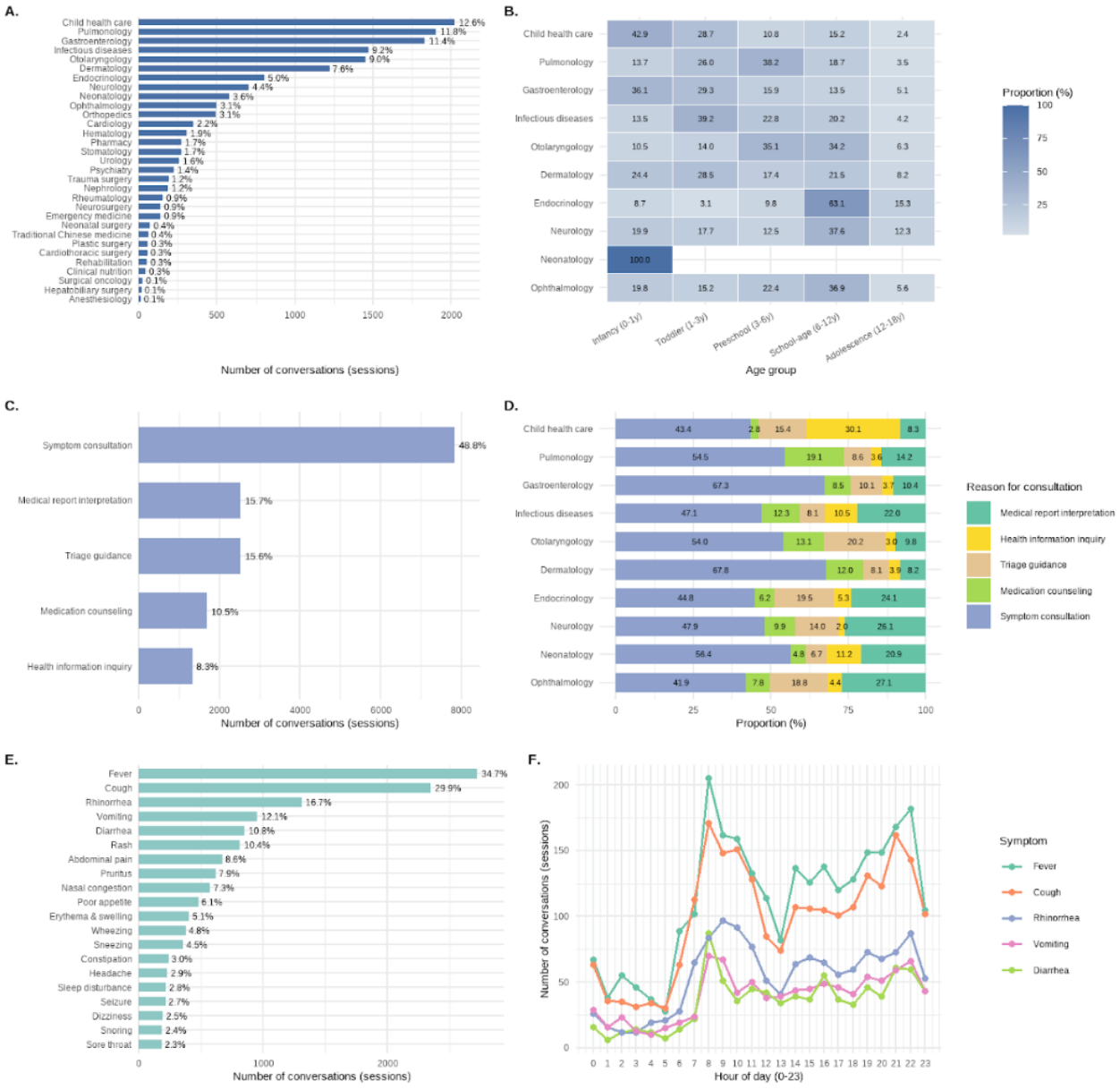
Reasons for consultation (Figures 3C and 3D) most frequently involved symptom consultation (7845; 48.8%), followed by medical report interpretation (2531; 15.7%), triage guidance (2515; 15.6%), medication counseling (1692; 10.5%), and health information inquiry (1333; 8.3%). The remaining 1% (n=158) were “other” queries about hospital services such as medical record copying, insurance, and campus locations. The distribution of consultation reasons also varied across subspecialties: medical report interpretation was most prevalent in ophthalmology (135/499, 27.1%) and neurology (184/705, 26.1%); health information inquiry in child health care (610/2025, 30.1%); triage guidance in otolaryngology (293/1454, 20.2%) and endocrinology (157/805, 19.5%);

XSL•FO
RenderX

medication counseling in pulmonology (364/1903, 19.1%); and symptom consultation in dermatology (827/1220, 67.8%) and gastroenterology (1234/1833, 67.3%).

Among the 7845 symptom consultation sessions, the most frequently reported symptoms were fever (2719; 34.7%), cough (2349; 29.9%), rhinorrhea (1309; 16.7%), vomiting (953; 12.1%), and diarrhea (848; 10.8%); the 20 most common symptoms are shown in Figure 3E, and all 43 standard symptoms are listed in Table S9 in Multimedia Appendix 1. When stratified by time of day, fever and cough were the 2 most frequently reported symptoms across daytime (8:00 to 18:00), evening (18:00 to 0:00), and overnight (0:00 to 8:00) periods, while their relative frequencies increased progressively from daytime to overnight (fever: 33.2% [1376/4144], 34.2% [881/2572], and 40.9% [462/1129]; cough: 28.4% [1176/4144], 29.9% [768/2572], and 35.9% [405/1129] of symptom consultation sessions within each period). The hourly distribution of the 5 most common symptoms across the 24-hour cycle is shown in Figure 3F.

Figure 3. Clinical consultation patterns across subspecialties, reasons, and symptoms. (A) Distribution of consultations across pediatric subspecialties. (B) Age distribution of consultations among the top 10 subspecialties. (C) Distribution of reasons for consultation. (D) Reasons for consultation across the top 10 subspecialties. (E) Top 20 symptoms reported by caregivers. (F) 24-hour distribution of top 5 reported symptoms.



Quality Evaluation Across Session and Turn Levels

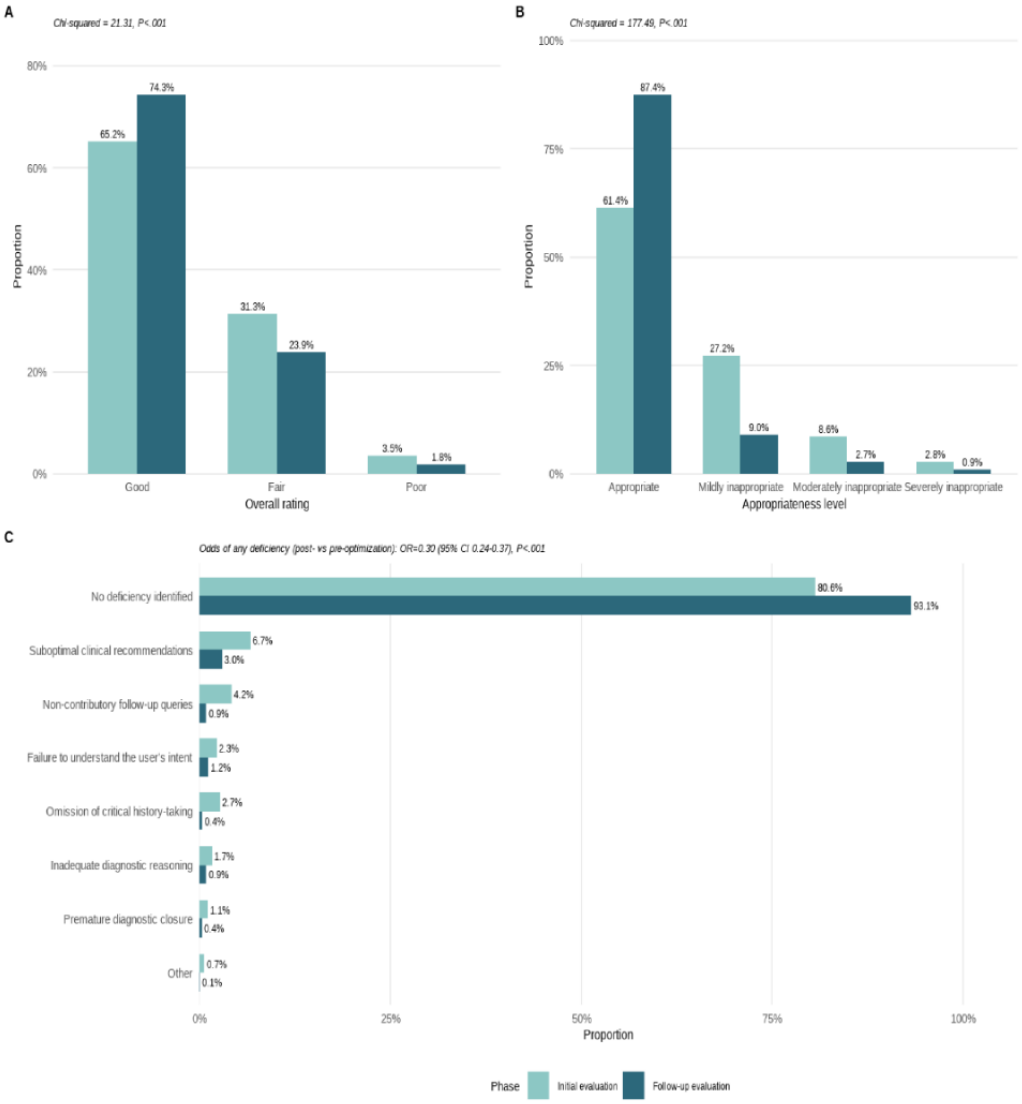
The initial and follow-up evaluations comprised 2000 sessions (1000 each) and 7492 turns (4073 and 3419, respectively). At the session level, the proportion rated good in overall quality increased from 65.2% (652/1000) to 74.3% (743/1000), and the proportion rated poor decreased from 3.5% (35/1000) to 1.8% (18/1000; χ^2_2=21.3; P<.001; Figure 4A). Clinical appropriateness also improved, with appropriate ratings rising from 61.4% (614/1000) to 87.4% (874/1000) and reductions across all 3 inappropriateness levels (χ^2_3=177.5; P<.001; Figure 4B). Sessions rated severely inappropriate decreased from 28

XSL•FO
RenderX

(2.8%) to 9 (0.9%), and those rated moderately inappropriate from 86 (8.6%) to 27 (2.7%).

At the turn level, the overall deficiency rate fell from 19.4% (790/4073) to 6.9% (236/3419; OR 0.30, 95% CI 0.24-0.37; *P*<.001), with declines across all 6 predefined deficiency categories and an "other" category (Figure 4C). Suboptimal clinical recommendations showed the largest reduction (from 6.7% [273/4073] to 3% [31/3419]) and remained the most frequent deficiency category after optimization: nonspecific specialty guidance (eg, suggesting pediatrics without a subspecialty) and medication advice lacking consideration of age or contraindications were largely resolved; remaining errors primarily concerned inadequate age-specific tailoring (eg, feeding volumes or dietary textures not adjusted for infant age). Noncontributory follow-up queries decreased from 4.2% (172/4073) to 0.9% (31/3419), with redundant or off-topic questioning largely eliminated; remaining errors mainly involved developmentally incongruent questions (eg, expecting a 1-year-old to self-report abdominal pain). Omission of critical history-taking decreased from 2.7% (112/4073) to 0.4% (14/3419). Fundamental gaps (eg, illness duration) were substantially reduced, with residual deficiencies confined to subspecialty-level depth. Failure to understand the user's intent decreased from 2.3% (94/4073) to 1.2% (41/3419), with residual errors primarily involving unaddressed caregiver concerns and unrecognized attempts at information correction. Inadequate diagnostic reasoning decreased from 1.7% (68/4073) to 0.9% (32/3419), with residual errors mainly comprising differential diagnosis for complex conditions. Premature diagnostic closure decreased from 1.1% (45/4073) to 0.4% (12/3419), and other deficiencies (0.7% [27/4073] to 0.1% [3/3419]; predominantly multimodal recognition and output issues) were effectively resolved.

Figure 4. Pediatrician evaluation of Xiaoyi responses across the two evaluations. (A) Overall quality rating in the initial and follow-up evaluations. (B) Response appropriateness in the initial and follow-up evaluations. (C) Turn-level response deficiencies in the initial and follow-up evaluations.



XSL•FO
RenderX

Discussion

Principal Findings

In this real-world analysis of a publicly deployed LLM-based pediatric health assistant, nearly half of consultations occurred outside routine clinic hours, and queries spanned a broad range of pediatric subspecialties and consultation content, suggesting potential use as a complement to traditional pediatric care. Expert-guided optimization yielded substantial gains in response quality, indicating a feasible pathway for refining deployed pediatric LLM systems. However, residual deficiencies persisted, underscoring the need for ongoing clinical oversight and refinement.

Consultations were sustained throughout the evening and overnight, with 43.8% (n=7046) occurring outside routine clinic hours. This pattern reflects a recognized mismatch between when caregivers seek guidance and when pediatric care is available. Parental health anxiety is common across populations and drives widespread online information-seeking for children's health [20]. This demand concentrates when traditional care is generally unavailable: pediatric telephone triage services report overnight peaks dominated by fever and respiratory symptoms [21], and an analysis of 500,000 health-related conversations on Microsoft Copilot found similar evening and overnight concentration [22]. Our findings are consistent with this broader pattern and extend it to pediatric AI: fever and cough were the most commonly reported symptoms, and the proportion of both rose progressively from daytime to overnight. Within the day, the peaks around 10:00 and 15:00 occurred during clinic hours and may reflect consultations initiated at or near the hospital or during breaks in the working day, whereas the peak around 21:00, within a sustained evening period of high volume, may coincide with children being at home, symptoms such as fever being noticed during bedtime routines, and outpatient services being closed. The similar weekday and weekend volumes may reflect our hospital's 7-day outpatient service, which leaves caregivers in its catchment area with no weekend gap in access. Compared with telephone triage, an AI assistant provides immediate and scalable responses unconstrained by staffing, an attribute of particular relevance in settings where pediatric workforce shortages and uneven geographic distribution of specialists are pronounced [23,24]. Taken together, these observations suggest that AI assistants may complement professional pediatric care when timely access is limited.

The breadth of consultation reasons observed in this study distinguishes our findings from prior pediatric AI studies, which have largely focused on single conditions or clinical scenarios. Symptom consultation constituted less than half of sessions, with medical report interpretation, triage guidance, medication counseling, and health information inquiry together forming a substantial remainder. The distribution of consultation reasons varied by subspecialty: report interpretation clustered in ophthalmology and neurology, specialties whose imaging-based reports and patient-facing materials have been shown to exceed the reading level recommended for laypeople [25,26]; health information queries concentrated in child health care, reflecting caregiver concerns about growth, development, and feeding [27]; and triage requests were most frequent in otolaryngology and endocrinology, areas in which referral thresholds and symptom urgency can be difficult to judge [28,29]. Collectively, these observations suggest that caregivers are using AI assistants as general-purpose pediatric advisors rather than single-purpose tools, and that the design and evaluation of future systems should account for this broad range of consultation content. These data also carry implications for the hospital itself: the concentration of report interpretation queries in ophthalmology and neurology identifies where existing patient-facing materials fall short, a gap that routine clinical encounters rarely reveal, and we are sharing these findings with the relevant departments to inform revision of these materials.

Our evaluation demonstrates that expert-guided optimization can yield quantifiable quality gains in a deployed pediatric LLM system. Across the 2 evaluations, the proportion of sessions rated as clinically appropriate rose from 61.4% (614/1000) to 87.4% (874/1000), and the turn-level deficiency rate declined from 19.4% (790/4073) to 6.9% (236/3419), with the largest reductions in suboptimal clinical recommendations and noncontributory follow-up queries. Clinically salient medication-safety errors, such as recommending chlorpheniramine for a 1.5-year-old or ibuprofen for an infant younger than 6 months, were no longer observed after optimization. This optimization pathway aligns with evidence that medication and guideline-related errors are more tractable when LLM outputs are grounded in curated knowledge sources and constrained by explicit safety checks [30,31]. The MEDIC system, for instance, reduced medication-direction near-miss events by extracting core prescribing elements, such as dose and frequency, and assembling final instructions using pharmacy logic and safety guardrails [32]; our combination of prompt revision, knowledge-base expansion, and validation rules reflects a similar approach.

However, this aggregate improvement does not imply that all errors are equally amenable to correction. Residual errors are concentrated in 2 more challenging categories: inadequate pediatric age- and developmental-stage tailoring, and deeper clinical reasoning involving subspecialty-level history-taking, caregiver concern recognition, and differential diagnosis for complex presentations. These patterns are consistent with evidence that current LLMs remain limited in pediatric contexts and in longitudinal, uncertainty-sensitive clinical reasoning, even when augmented by prompting or retrieval-based strategies [33,34]. Pediatric underrepresentation in AI and LLM research may contribute to the former [35,36], whereas the latter reflects the persistent difficulty of multistep clinical reasoning and differential diagnosis for current LLMs [37]. Addressing these residual errors may require pediatric-specific fine-tuning, dedicated reasoning modules, or human-in-the-loop referral for high-complexity cases.

The clinical significance of these residual deficiencies is better reflected by the session-level clinical appropriateness rating, which is graded according to the potential for harm, than by the turn-level deficiency rate, which counts any deficiency regardless of severity. After optimization, most sessions not rated appropriate were mildly inappropriate (90/1000), indicating minor issues that do not affect health judgment or family

XSL•FO
RenderX

decision-making, and the residual turn-level deficiencies were predominantly deficits of specificity rather than incorrect guidance. Nonetheless, 9 sessions (0.9%) remained severely inappropriate, a proportion that is not negligible in a service used by caregivers without clinical training and that warrants continued safety monitoring.

Xiaoyi does not currently elicit an environmental history, although environmental exposures contribute to many of the respiratory, allergic, and skin complaints raised by caregivers [5], and none of our deficiency categories captured this omission. Structured instruments such as the WHO green page provide a framework for the pediatric environmental history, and self-completed digital versions have been implemented in mobile health interventions [38]. A brief, complaint-triggered set of environmental questions, together with environmental data such as air quality indices and pollen counts in the regional epidemiology tool, represents a practical direction for future development.

Transferability to other countries rests on the system's architecture and development cycle rather than on the deployed system itself: retrieval-augmented reasoning with deterministic clinical tools and a postgeneration validation layer, refined through iterative expert evaluation. Several components are locally specific. The reasoning core was selected for its Chinese-language performance and would require replacement rather than translation; the knowledge bases and medication rules follow Chinese guidelines and the national pediatric formulary; and the triage logic presumes a children's hospital organized by subspecialty. Data protection requirements and the jurisdiction-dependent boundary between health guidance and regulated medical software impose further constraints, and transferability outside China remains to be tested.

This study has several limitations. First, although the platform reached caregivers from 31 of 34 provincial-level administrative divisions in China, registrants were predominantly from Chongqing and adjacent regions, reflecting the sponsoring institution's catchment area. Patterns of use may differ in settings with other health care infrastructure, digital connectivity, or caregiver familiarity with AI-based services, and external validation in additional regions will be needed to establish broader generalizability. Promotional activities may also have influenced the composition of the user base and consultation patterns during those periods, and IP-based geographic inference cannot detect the use of virtual private networks or proxy servers. Second, each session was assigned to a single subspecialty-matched rater, so formal interrater reliability was not estimated. The same expert panel also rated sessions across both evaluation rounds, consistent with the study's quality-improvement design but without phase blinding. Future evaluations incorporating multirater scoring and blinded assessment would provide additional methodological reinforcement. Third, some evaluation criteria were setting-specific: classifying referral to general pediatrics as nonspecific suits a tertiary children's hospital organized by subspecialty and may not transfer to systems in which general pediatricians provide first-contact care. In addition, safety was assessed through the graded clinical appropriateness rating of sampled sessions rather than through independent annotation of critical safety errors across all interactions. Fourth, within this observational design, the contribution of expert-guided optimization cannot be fully isolated from that of ongoing platform refinements; the consistency of improvement across deficiency categories and its timing relative to the optimization phase; however, support a predominant role of expert guidance. Fifth, although each child was assigned a unique profile and caregiver accounts were device-independent, a child registered separately by 2 caregivers could not be identified from the de-identified data; correlation among sessions from the same child was therefore not modeled. Sixth, the symptom lexicon reflects the expressions caregivers used during the study period and may under-represent symptom categories that were infrequently reported; prospective expansion of the lexicon would improve sensitivity for environmentally related presentations. Finally, our evaluation focused on pediatrician-rated response quality rather than on downstream clinical outcomes, including appropriate triage to in-person care, diagnostic concordance, symptom resolution, and adverse events. Future work should examine whether such assistants change care-seeking behavior, including whether they reduce unnecessary visits or increase the proportion of consultations in which in-person assessment is warranted, and should survey local pediatricians on their perceptions of the tool and on any resulting changes in clinical practice and family expectations.

Conclusions

In conclusion, this real-world analysis of a publicly deployed LLM-based pediatric health assistant showed that a substantial proportion of caregiver consultations occurred outside routine clinic hours, and queries spanned a broad range of pediatric subspecialties and clinical concerns. Expert-guided optimization was associated with measurable improvements in pediatrician-rated response quality, while exposing persistent challenges in age- and developmental-stage-specific assessment and higher-order clinical reasoning. These findings support cautious optimism for LLM-based assistants as a scalable adjunct to pediatric care, while underscoring the need for sustained expert oversight, pediatric-specific safety frameworks, and prospective evaluation of clinical outcomes as such systems advance toward broader implementation in pediatric clinical practice.

Acknowledgments

We are grateful to all the caregivers who participated in this research. We sincerely thank the following pediatricians from Children's Hospital of Chongqing Medical University, who served on the expert evaluation panel and contributed their clinical expertise and valuable time to the evaluation of Xiaoyi's performance: Yongmei Li, Xin Xiang, Feng Zhang, Zhenzhen Zhang, Hongqiang Du, Wenjing Tang, Xuemei Tang, Cheng Zhong, Sijie Yu, Anshuo Wang, Long Chen, Yan Shu, Lei Zhang, Xue Zhou, Huichao Sun, Xia Li, Yaping Deng, Chunlan Qiu, Yi Wang, Jinping Hou, Liangqiao Tao, Deying Zhang, Na Zang, Jinwei

XSL•FO
RenderX

Tu, Yingzhu Kang, Ying Xu, Zhu Li, Feng Chen, Lijuan Ouyang, Wenquan Cai, Yonggang Li, Yan Xue, Xiao Fang, and Zehui Ye.

Funding

This study was supported by the Natural Science Foundation of Chongqing, China (CSTB2025NSCQ-LZX0062), the China Postdoctoral Science Foundation (2024MD764064), and the Special Funding for Postdoctoral Research Project of Chongqing (2024CQBSHTB3014). The funders had no role in study design, data collection, analysis, interpretation, or writing of the manuscript.

Data Availability

The dataset analyzed during this study is not publicly available due to restrictions in the ethics approval but may be available from the corresponding author upon reasonable request. Further technical details of the Xiaoyi system, including model training and deployment specifications, are also available from the corresponding author upon reasonable request.

Generative AI Use

Generative AI (DeepSeek-V3.2) was used solely for language editing to improve the grammar, clarity, and readability of author-written text. It was not used for study design, data analysis or interpretation, or the generation of scientific content or references. All suggested edits were reviewed and, where necessary, revised by the authors, who take full responsibility for the content of this manuscript.

Authors' Contributions

Conceptualization: XX, ZH, PS
Formal analysis: QW, LX, XZ
Funding acquisition: XX, QW
Investigation: LX, QW, HW, RW, SO, ZH
Software: CZ, HH, YP, LL
Methodology: QW, XZ, XX
Writing original draft: QW
Writing review & editing: XX, XZ, LX
Supervision: XX, PS, ZH
Project administration: ZH, PS
QW, LX, and XZ contributed equally to this work and should be considered joint first authors. XX (ximing@hospital.cqmu.edu.cn), ZH (zyhua@cqmu.edu.cn), and PS (pingsong@hospital.cqmu.edu.cn) contributed equally to this work and should be considered joint corresponding authors.

Conflicts of Interest

CZ is the CEO of Zuoyi Tech, and HH, YP, and LL are employees of Zuoyi Tech. Zuoyi Tech group participated in the technical development of Xiaoyi. The other authors declare no competing interests.

Multimedia Appendix 1

Methods, annotation prompts, optimization strategies, assessment criteria, and consultation characteristics.
[DOCX File , 1415 KB-Multimedia Appendix 1]

References

1. Pöyry H, Turunen J, Ritola E, Hartikainen S, Palviainen J, Liimatta I, et al. Parental ability to identify severe illnesses in their children. JAMA Netw Open. 2026;9(2):e2559998. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.59998] [Medline: 41701491]
2. Russ C, Gao Y, Karpowicz K, Lee S, Stephens TN, Trimm F, et al. The pediatrician workforce in the United States and China. Pediatrics. Jun 01, 2023;151(6):e2022059143. [doi: 10.1542/peds.2022-059143] [Medline: 37158018]
3. Westin E, Svensson A, Gustafsson IL, Elmqvist C, Levander MS. Parents' experiences of telephone nursing when their child has a fever: a mixed method study. BMC Nurs. 2026;25(1):339. [FREE Full text] [doi: 10.1186/s12912-026-04585-0] [Medline: 41913143]
4. Borch-Johnsen L, Folke F, Frederiksen MS, Schrøder M, Greisen G, Lund S, et al. Parents' motivations for calling an out-of-hours helpline: qualitative study. J Med Internet Res. 2025;27:e66780. [FREE Full text] [doi: 10.2196/66780] [Medline: 40373297]

XSL•FO
RenderX

5. Campillo I López F. It is time for planetary paediatrics. An Pediatr (Engl Ed). 2022;96(3):177-178. [FREE Full text] [doi: 10.1016/j.anpede.2021.11.004] [Medline: 35241405]
6. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
7. Wan P, Huang Z, Tang W, Nie Y, Pei D, Deng S, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. Nat Med. 2024;30(10):2878-2885. [doi: 10.1038/s41591-024-03148-7] [Medline: 39009780]
8. Tao X, Zhou S, Ding K, Li S, Li Y, Wu B, et al. An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nat Med. Mar 2026;32(3):934-942. [doi: 10.1038/s41591-025-04176-7] [Medline: 41555035]
9. Sun Y, Xu S, Jin H, Han X, Jin K, Zhang Y, et al. Effectiveness of Al-assisted patient health education using voice cloning and ChatGPT: prospective randomized controlled trial. J Med Internet Res. 2026;28:e81387. [FREE Full text] [doi: 10.2196/81387] [Medline: 41855493]
10. Holderried F, Sonanini A, Stegemann-Philipps C, Herrmann-Werner A, Spitzer P, Guthoff M, et al. Impact of GPT-4-generated discharge letters on patients' medical comprehension: prospective crossover study. J Med Internet Res. 2026;28:e81243. [FREE Full text] [doi: 10.2196/81243] [Medline: 41746691]
11. Zhao Y, Qian W, Chen Y, Wu D, Luo Y, Gao C, et al. Effect of an AI agent trained on a large language model (LLM) as an intervention for depression and anxiety symptoms in young adults: A 28-day randomized controlled trial. Appl Psychol Health Well Being. Oct 2025;17(5):e70067. [doi: 10.1111/aphw.70067] [Medline: 40910958]
12. Callahan C, Tanner L, Coe C, Davis M, Glover J, Bernstein E, et al. Real-world use of a mental health AI companion: multiple methods study. JMIR Form Res. 2026;10:e86904. [FREE Full text] [doi: 10.2196/86904] [Medline: 41687100]
13. Barile J, Margolis A, Cason G, Kim R, Kalash S, Tchaconas A, et al. Diagnostic accuracy of a large language model in pediatric case studies. JAMA Pediatr. 2024;178(3):313-315. [doi: 10.1001/jamapediatrics.2023.5750] [Medline: 38165685]
14. Mansoor M, Ibrahim A, Hamide A. Performance of DeepSeek and GPT models on pediatric board preparation questions: comparative evaluation. JMIR AI. 2025;4:e76056. [FREE Full text] [doi: 10.2196/76056] [Medline: 40864882]
15. Wei Q, Wang Y, Yao Z, Cui Y, Wei B, Li T, et al. Evaluation of ChatGPT's performance in providing treatment recommendations for pediatric diseases. Pediatr Discov. 2023;1(3):e42. [doi: 10.1002/pdi3.42] [Medline: 40626087]
16. Chen SF, Alyakin A, Seas A, Yang E, Choi JJ, Lee JV, et al. LLM-assisted systematic review of large language models in clinical medicine. Nat Med. 2026;32(3):1152-1159. [doi: 10.1038/s41591-026-04229-5] [Medline: 41776077]
17. Ouyang H, Lin G, Li Y, Yao Z, Li Y, Yan H, et al. Evaluating and validating large language models for health education on developmental dysplasia of the hip: 2-phase study with expert ratings and a pilot randomized controlled trial. J Med Internet Res. 2026;28:e73326. [FREE Full text] [doi: 10.2196/73326] [Medline: 41554120]
18. Aydemir U. ChatGPT-delivered physical activity intervention for children with autism spectrum disorder: pre-post feasibility study. JMIR Hum Factors. 2025;12:e71119. [FREE Full text] [doi: 10.2196/71119] [Medline: 40498657]
19. CHART Collaborative, Huo B, Collins GS, Chartash D, Thirunavukarasu AJ, Flanagin A, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open. 2025;8(8):e2530220. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.30220] [Medline: 40747871]
20. Kubb C, Foran HM. Online health information seeking by parents for their children: systematic review and agenda for further research. J Med Internet Res. 2020;22(8):e19985. [FREE Full text] [doi: 10.2196/19985] [Medline: 32840484]
21. Lous J, Moth G, Huibers L, Vedsted P, Christensen MB. Preschool children in Danish out-of-hours primary care: a one-year descriptive study of face-to-face consultations. BMC Fam Pract. 2019;20(1):36. [FREE Full text] [doi: 10.1186/s12875-019-0922-y] [Medline: 30808295]
22. Costa-Gomes B, Tolmachev P, Taysom E, Sounderajah V, Richardson H, Schoenegger P, et al. Public use of a generalist LLM chatbot for health queries. Nat. Health. Apr 16, 2026;1(7):689-696. [doi: 10.1038/s44360-026-00117-x]
23. Liu Y, Yang LL, Xu SY, Zhao ZY. Pediatrics in China: challenges and prospects. World J Pediatr. 2018;14(1):1-3. [doi: 10.1007/s12519-017-0117-3] [Medline: 29383583]
24. Turner A, Ricketts T, Leslie LK. Comparison of number and geographic distribution of pediatric subspecialists and patient proximity to specialized care in the US Between 2003 and 2019. JAMA Pediatr. Sep 01, 2020;174(9):852-860. [FREE Full text] [doi: 10.1001/jamapediatrics.2020.1124] [Medline: 32421165]
25. Dihan QA, Brown AD, Alzein AF, Zaldivar AT, Montgomery KE, Khodeiry MM, et al. Enhancing patient and parent education in pediatric ophthalmology using artificial intelligence: a report by the AAPOS public information committee. J AAPOS. Dec 2025;29(6):104693. [doi: 10.1016/j.jaapos.2025.104693] [Medline: 41274543]
26. Alabed S, Anderson A, Maiter A, Hughes A, McAnenly N, Salehi M, et al. Large language models for simplifying radiology reports: a systematic review and meta-analysis of patient, public, and clinician evaluations. Lancet Digit Health. Feb 2026;8(2):100960. [FREE Full text] [doi: 10.1016/j.landig.2025.100960] [Medline: 41698858]
27. De Rosso S, Nicklaus S, Ducrot P, Schwartz C. Information seeking of French parents regarding infant and young child feeding: practices, needs and determinants. Public Health Nutr. 2022;25(4):879-892. [FREE Full text] [doi: 10.1017/S1368980021003086] [Medline: 34321131]

XSL•FO
RenderX

28. Algadi MA, Alshathri AA, Alsugair RS, Alyabis MA, Alsaleh SA, Aljerian NA. Trends and patterns in urgent pediatric otolaryngology inter-hospital referrals in Saudi Arabia. Saudi Med J. 2022;43(1):91-97. [FREE Full text] [doi: 10.15537/smj.2022.43.1.20210710] [Medline: 35022289]
29. Nakhla M, Cuthbertson D, Becker DJ, Pacaud D, Ludvigsson J, Knip M, et al. Diabetic ketoacidosis at the time of diagnosis of type 1 diabetes in children: insights from TRIGR. JAMA Pediatr. 2021;175(5):518-520. [FREE Full text] [doi: 10.1001/jamapediatrics.2020.5512] [Medline: 33492389]
30. Gilbert S, Kather JN, Hogan A. Augmented non-hallucinating large language models as medical information curators. NPJ Digit Med. 2024;7(1):100. [FREE Full text] [doi: 10.1038/s41746-024-01081-0] [Medline: 38654142]
31. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. [FREE Full text] [doi: 10.1038/s41746-025-01519-z] [Medline: 40185842]
32. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M. Large language models for preventing medication direction errors in online pharmacies. Nat Med. 2024;30(6):1574-1582. [FREE Full text] [doi: 10.1038/s41591-024-02933-8] [Medline: 38664535]
33. Huang T, Tse G, Pageler N, Bannett Y. Large language models using clinical text in pediatrics: a scoping review. JAMA Netw Open. 2026;9(3):e262443. [FREE Full text] [doi: 10.1001/jamanetworkopen.2026.2443] [Medline: 41879783]
34. 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):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
35. Richter F, Holmes E, Richter F, Guttmann K, Duong SQ, Gangadharan S, et al. Toward governance of artificial intelligence in pediatric healthcare. NPJ Digit Med. 2025;8(1):636. [FREE Full text] [doi: 10.1038/s41746-025-02000-7] [Medline: 41168479]
36. Muralidharan V, Burgart A, Daneshjou R, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. NPJ Digit Med. 2023;6(1):166. [FREE Full text] [doi: 10.1038/s41746-023-00898-5] [Medline: 37673925]
37. Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, et al. Large language model performance and clinical reasoning tasks. JAMA Netw Open. 2026;9(4):e264003. [FREE Full text] [doi: 10.1001/jamanetworkopen.2026.4003] [Medline: 41973425]
38. Ortega-García JA, Sánchez-Sauco MF, Zafra-Rodríguez JA, Cabrera-Rivera LT, Díaz-Martínez F, Llegus-Santiago EM, et al. Subjective well-being, happiness, and environmental health factors related to women planning a pregnancy or pregnant, using mobile health intervention. Digit Health. 2023;9:20552076231177146. [FREE Full text] [doi: 10.1177/20552076231177146] [Medline: 37284011]

Abbreviations

CHART: Chatbot Assessment Reporting Tool
GEE: generalized estimating equations
LLM: large language model
OR: odds ratio
ReAct: reasoning-and-acting
WHO: World Health Organization

Edited by M Balcarras; submitted 24.Jun.2026; peer-reviewed by F Campillo, E Reiter; comments to author 17.Sep.2026; revised version received 11.Oct.2026; accepted 13.Oct.2026; published 04.Nov.2026

Please cite as:
Wei Q, Xiao L, Zhang X, Wei H, Wang R, Ou S, Zhang C, Huo H, Peng Y, Liu L, Song P, Hua Z, Xu X
Real-World Use and Expert-Guided Optimization of a Large Language Model–Based Pediatric Health Assistant: Retrospective Observational Study
J Med Internet Res 2026;28:e105346
URL: https://www.jmir.org/2026/1/e105346
doi: 10.2196/105346
PMID:

©Qiuhong Wei, Li Xiao, Xin Zhang, Hong Wei, Ruijue Wang, Suqiao Ou, Chao Zhang, Huarong Huo, Yu Peng, Lishen Liu, Ping Song, Ziyu Hua, Ximing Xu. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 04.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License

XSL•FO
RenderX

(https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX