frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 22 September 2025
DOI 10.3389/frai.2025.1640994

OPEN ACCESS

EDITED BY
Anna Sandionigi,
Quantia Consulting srl, Italy

REVIEWED BY
Youssef Er-Rays,
Abdelmalek Essaadi University, Morocco
Sarita Agrawal,
All India Institute of Medical Sciences, Raipur, India

*CORRESPONDENCE
Pulidindi Venugopal
✉ pulidindi.venu@vit.ac.in

RECEIVED 04 June 2025
ACCEPTED 22 August 2025
PUBLISHED 22 September 2025

CITATION
Valan P and Venugopal P (2025) Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility–accuracy-readability study of the DIAN AI assistant.
Front. Artif. Intell. 8:1640994.
doi: 10.3389/frai.2025.1640994

Evaluating a retrieval-augmented pregnancy chatbot: a comprehensibility–accuracy-readability study of the DIAN AI assistant

P. Valan and Pulidindi Venugopal*

VIT Business School, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Introduction: Patient education materials (PEMs) often exceed common health literacy levels. Retrieval-augmented conversational AI may deliver interactive, evidence-grounded explanations tailored to user needs. We evaluated DIAN, a RAG-enabled pregnancy chatbot grounded in the NHS Pregnancy Book, using a comprehensibility–accuracy–readability (CAR) framework to compare perceptions between women and clinicians across key perinatal domains.

Methods: We conducted a cross-sectional evaluation with standardized prompts and blinded scoring. Participants were 119 women (18–55 years) and 29 clinicians. After brief CAR training and calibration, all evaluators independently rated the same DIAN responses on 4-point Likert scales across postpartum care, pregnancy health and complications, diet and nutrition, and mental and emotional wellbeing. Between-group differences were tested using the Mann–Whitney U test with Bonferroni adjustment across domains per outcome; effect sizes were summarized with $r = |Z|/\sqrt{N}$ and Cliff's delta. Inter-rater reliability was not estimated, given the independent-rater design.

Results: Differences concentrated in postpartum care. Comprehensibility favored women ($U = 1206.50$, $Z = -2.524$, $p = 0.012$; $r = 0.207$; $\Delta = 0.301$). Accuracy also favored women ($U = 1239.00$, $Z = -2.370$, $p = 0.018$; $r = 0.195$; $\Delta = 0.282$). Readability favored clinicians ($U = 1181.50$, $Z = -2.639$, $p = 0.008$; $r = 0.217$; $\Delta = 0.315$). Other domains showed no significant between-group differences after correction. Radar visualizations mirrored these patterns, with women showing larger comprehensibility/accuracy profiles and clinicians showing larger readability profiles in postpartum care.

Discussion: Grounded in an authoritative national guide, DIAN achieved broadly comparable CAR perceptions across groups, with clinically relevant divergence limited to postpartum care. Women perceived higher comprehensibility and accuracy, while clinicians judged language more readable, suggesting a gap between experiential clarity and professional textual ease. Targeted postpartum refinement, lexical simplification, role-tailored summaries, and actionable checklists may align perceptions without compromising fidelity. More broadly, RAG-grounded chatbots can support equitable digital health education when content is vetted, updated, and evaluated with stakeholder-centered metrics. Future work should examine free-form interactions, longitudinal behavioral outcomes, and ethical safeguards (scope-of-use messaging, escalation pathways, and bias audits).

KEYWORDS

patient education material, health communication, maternal health, chatbot, patient education

1 Introduction

Patient education is fundamental to quality healthcare, yet significant barriers persist in delivering accessible and comprehensible health information to patients and their families. Online search engines are a common source of health information for patients; however, the reliability and accuracy of these digital resources are questionable (Abdullah et al., 2022). This challenge is particularly pronounced with patient education materials (PEMs), which are educational resources designed to inform patients about their disease or illness (Saunders et al., 2018). Research consistently demonstrates that many PEMs are too complex for patients with less than a high school education (Nattam et al., 2023), creating a critical need to enhance the accessibility and usability of high-quality materials for patients and their families (Slatore et al., 2016). To overcome these barriers and ensure patient understanding, innovative technological solutions are needed.

The aim of this research is to evaluate whether an AI-powered conversational agent, equipped with Retrieval-Augmented Generation (RAG), can improve the comprehensibility, accuracy, and readability of patient education content for diverse user groups, specifically women and healthcare professionals in obstetric care. To address the limitations of current PEMs in accuracy and readability, we consider how advances in AI can directly support clearer, more accessible patient communication.

Artificial intelligence (AI) presents unprecedented opportunities to address these patient education challenges. The implementation of AI in healthcare provides detailed technical support through computer intelligence combined with human intervention (Du et al., 2019). However, successful AI implementation in the medical field requires two components: structured data for machine learning and regular training data to enhance system performance (Jiang et al., 2017). The growing nature of data in healthcare leads to AI implementation faster and covers different verticals of healthcare (Ganesh et al., 2022). The inclusion of AI in healthcare has been found to be supportive in communicable or non-communicable disease diagnostics, assessing the risk of mortality and morbidity, predicting and surveilling the outbreak of diseases, and planning and drafting health policies (Schwalbe and Wahl, 2020). While AI offers many benefits across healthcare, its true impact on patient education is realized through more interactive and personalized tools. Within AI applications, conversational AI specifically targets the comprehension and engagement gaps of static PEMs by enabling tailored, interactive explanations.

The word "conversation" may mislead us into understanding conversational AI (ConAI) as an enterprising system with the ability to converse (Elbanna, 2007), but ConAI's potential to represent human competence, ability to learn and improvise (Gkinko and Elbanna, 2023), and possess a personality trait (Bavaresco et al., 2020) makes it unique among AI models. The inclusion of ConAI will result in focused customization, leading to anthropomorphism (Fotheringham and Wiles, 2023), and creates an opportunity for more engaging and effective patient education experiences (Speirs, 2015). With the help of ConAI, patients can interact with PEMs in real time and receive information in understandable, context-sensitive language. This interactive approach helps bridge knowledge gaps and reduces anxiety around medical care (Akpan et al., 2025). By simplifying complex medical jargon, ConAI makes educational materials more accessible to patients with varying literacy levels (Nasra et al., 2025). Gathering routine patient queries, educating about medications, procedures, and aftercare, and collecting preliminary information before treatments enable clinicians to focus on more complex tasks and improve workflow efficiency (Hong et al., 2022); thus, ConAI becomes a clinical companion. Using ConAI, we can achieve broad reach demands for public health campaigns, such as vaccinations or chronic disease management, and can deliver patient education on a large scale, making high-quality information available to underserved or remote populations (Mondal et al., 2023). Recent innovations in ConAI include "synthetic patients" AI-driven avatars that simulate challenging patient conversations. These tools are now used for training medical students in soft skills such as delivering bad news, managing emotional responses, and addressing health literacy gaps (Chu and Goodell, 2024). However, the effectiveness of conversational agents depends not only on their language abilities and interactivity but also on the accuracy and currency of the information they provide. Yet, despite these advantages, conversational agents can suffer from inaccuracies, outdated content, and inconsistent intent-matching limitations that necessitate explicit grounding in authoritative evidence.

However, to ensure that conversational outputs are not only engaging but also accurate and current, Retrieval-Augmented Generation (RAG) provides a mechanism to ground responses in trustworthy sources.

The RAG addresses these limitations by coupling retrieval of vetted sources with generation, thereby improving factual correctness, currency, and transparency in patient-facing explanations. In some variants, the retriever and generator are trained end-to-end, retrieving evidence passages via dense question–passage similarity to enhance accuracy and interpretability (Lewis et al., 2020; Oche et al., 2025); this grounding is especially valuable for patient education, where conversations must rely on trustworthy, evidence-based sources.

In general, RAG is a method that helps large language models (LLMs) give better answers by first searching for and using information from outside sources, such as documents (Tian et al., 2025). This capability is especially valuable in the context of patient education, where conversations must draw from trustworthy, evidence-based sources to deliver safe and effective information.

In patient education, ensuring information reliability is paramount. RAG models can cite validated, up-to-date medical literature or institutional guidelines in real time, supporting patient safety and regulatory compliance (Gargari and Habibi, 2025). Using RAG to improve the accuracy, reliability, and specificity of clinical responses, especially in knowledge-intensive medical tasks, can be made easy (Shin et al., 2025). "Medicare" and other medical dialog settings will be leveraging RAG (Agrawal1 et al., 2025) for shared medical decision-making (Shi et al., 2023).

Despite promising advances in conversational AI and the integration of retrieval-augmented models, there remains a lack of empirical evidence regarding their effectiveness in improving patient education comprehensibility, accuracy, and readability, especially across different user groups. This research is expected to demonstrate that leveraging advanced conversational AI can enhance the accessibility and clarity of patient education materials, particularly for lay audiences, offer practical insights into tailoring digital health content for distinct user groups using structured evaluation frameworks, and inform future development and deployment strategies for digital patient education solutions across healthcare settings.

2 Related works

To contextualize our study, we examined previous research on healthcare chatbots addressing various scenarios, including maternal and reproductive health. Table 1 highlights examples of chatbots developed for different purposes, including fertility awareness (Maeda et al., 2020), gestational diabetes management (Sagstad et al., 2022), maternal health (Nguyen et al., 2024), and perinatal mental health (Chung et al., 2021). Table 2 details their methodologies, showcasing the advancements and study aims addressed. Most existing studies have either focused on technical performance metrics or limited use cases within narrow clinical contexts. Few have systematically compared the perceptions and understanding of both healthcare professionals and lay audiences, gaps that are critical to address as such technologies become increasingly integrated into patient care. Prior work on healthcare chatbots illustrates promise and common limitations, motivating a closer look at how users actually perceive comprehensibility, accuracy, and readability in real use. Building on these considerations, we focus our study on how different user groups perceive the same chatbot content across key obstetric domains.

We examine comprehensibility–accuracy–readability (CAR) for a RAG-enabled obstetric chatbot among women and doctors across postpartum care, pregnancy health and complications, prenatal preparation and support, diet and nutrition, mental and emotional wellbeing, birth preferences and experiences, and practical preparations for baby.

1. Does the use of a RAG-powered conversational chatbot improve the comprehensibility, accuracy, and readability of patient education materials compared to standard resources, as evaluated by both women (patients) and healthcare professionals?

TABLE 1 Chatbots developed for patient education and health promotion.

Name of chatbot	Fertility chatbot	Dina	Rosie	Dr. Joy
Purpose	Promote fertility awareness and preconception health	Supports pregnant women with GDM by providing guidance	Provides personalized health education for new mothers	Provides obstetric and mental health care support
Target audience	Women aged 20 to 34 years	Pregnant women with gestational diabetes in Norway	Mothers of color, currently pregnant or with infants <6 months	Perinatal women and their partners in South Korea
Platform	Online website	Online and Norwegian digital health platform	Mobile app (iPhone and Android)	Mobile instant messaging app (KakaoTalk)
Development approach	Pre-programmed scripted chatbot	User-centered design with health expert input	Community-driven design over 3 years	Text-mining with input from 11 medical specialists
Core features	Fertility education, RLP counseling, and knowledge improvement	Blood glucose management, diet tips, and physical activity advice	Parenting advice, child development, and health emergency detection	Obstetric and mental health Q&A, symptom checks, CBT tools
Health focus	Fertility and preconception health	Gestational Diabetes Mellitus (GDM)	Maternal and child health	Obstetrics and mental health
Evaluation method	Three-arm randomized controlled trial	Observational study analyzing chatbot logs	Randomized pilot study with treatment and control groups	Usability testing (7-day contextual study)
Key findings	Improved fertility knowledge and behavior intentions, reduced anxiety	Answered 88.51% of questions, mirrored GDM treatment priorities	Reduced postpartum depression, improved health info accessibility	High usability, positive associations with perceived benefits
Limitations	Technical limitations, low comprehension of user inputs	Limited content scope, need for promotion	App crashes, user dissatisfaction with some responses	Limited intent matching, content requires regular updates
Usability features	Simplified educational content, feedback integration	Low-threshold information access, available anytime	Daily push notifications, video tutorials, FAQs	Multi-language support, dialog buttons for guided flow
Personalization level	Moderate (pre-determined scenarios)	Moderate (focus on GDM-related questions)	High (personalized based on user queries)	Moderate to high (rich knowledge base and synonym dictionary)
Accessibility	Accessible via website	Freely available without registration	Accessible on mobile devices	Available via popular instant messenger
Integration with existing systems	Standalone system without broader integration	Integrated with the national digital health platform	Limited integration; standalone app	Integrated within the KakaoTalk platform
Primary outcomes	Increased fertility knowledge, behavior intentions without anxiety increase	Improved access to GDM-related information	Statistically significant reduction in postpartum depression	Enhanced user engagement with medical Q&A and mental health tools

TABLE 2 Previously Used Chatbot Development Methods for Patient Education and Health Promotion.

Name of chatbot	Unnamed fertility chatbot	Dina	Rosie	Dr. Joy
Data used	Pre-designed fertility and preconception health education scripts from trusted sources like the Japan Society of Obstetrics and Gynecology.	Anonymous dialog logs from 610 interactions categorized into themes such as glucose management, diet, and physical activity.	Community feedback through focus groups, listening sessions, and 73,000 expert-vetted passages from trusted medical sources.	3,524 refined Q&A pairs from South Korea's largest online maternal care community, enhanced with synonyms and neologisms.
Development methodology	Scripted chatbot with predetermined conversational scenarios; content simplified for readability and casual interaction.	User-centered design with input from clinicians, focusing on self-management education aligned with national GDM guidelines.	Community-driven iterative development over 3 years; built a knowledge base with FAQs and push notifications tailored to user needs.	Developed using KakaoTalk's AI platform with input from 11 medical specialists, focusing on conversational responses and emotional warmth.
Testing methodology	Three-arm randomized controlled trial with 927 women divided into intervention and control groups, evaluating knowledge, behavior intentions, and anxiety.	Observational study analyzing chatbot logs over two time periods, monitoring query categories and fallback rates.	Randomized pilot study with 29 participants, split into chatbot and control groups, evaluating usage, feedback, postpartum depression, and emergency room visits.	7-day contextual usability test with 15 participants providing feedback on daily chatbot interactions, tracked using emojis and qualitative comments.
Results	Improved fertility knowledge and behavior intentions; reduced anxiety; highlighted the need for better user comprehension capabilities.	Answered 88.51% of questions; reflected GDM treatment priorities; recommended better content and wider promotion.	Significant reduction in postpartum depression; high usability but technical issues like app crashes; recommendations for refining responses and app stability.	High usability and positive benefits; appreciated professional content but noted intent matching limitations and need for periodic updates.

2. Are there discernible differences in these perceptions (CAR scores) between lay users and clinicians across common patient education domains such as postpartum care, pregnancy health, diet and nutrition, and mental and emotional wellbeing?
3. What are the implications for personalized and effective digital health education for diverse user groups?

3 Materials and methods

3.1 Participant recruitment and study setting

A multi-modal recruitment strategy was employed to recruit study participants. Healthcare providers at obstetric clinics were approached and provided with study information materials to facilitate recruitment (Rokicki et al., 2025). Recruitment advertisements were placed within participating clinics to maximize visibility among the target population. Potential participant details were collected, and the research team contacted them for further screening and selection.

The inclusion criteria specify that women participants must be 18 years of age or older. Women who were currently pregnant, had experience in pregnancy in the past, or had closely supported someone during pregnancy were considered participants. Irrespective of their educational background, participants were included only if they were able to read and communicate in English, as all study materials and interviews were conducted in English. Women who have back ground in medicine and any connections with healthcare professionals were excluded to ensure perspectives reflected lay experiences.

Healthcare professionals (doctors) were eligible if they were licensed medical doctors (MBBS or equivalent), had at least 2 years of clinical experience in maternal or prenatal healthcare, were currently practicing or had practiced within the past 5 years, were fluent in English, and provided informed consent. Doctors were excluded if they had previous involvement in developing the AI chatbot or any related study.

We then selected two groups of participants to test the DIAN chatbot: 29 healthcare professionals with substantial years of experience and 119 women.

3.2 Identification of key concerns through thematic analysis

We gathered queries from research participants to comprehensively understand the recurring questions and concerns encountered by them. During semi-structured interviews, participants were prompted with open-ended questions such as, "What are some common questions or concerns that come up during pregnancy?," "Are you aware of any patient education materials related to pregnancy?," and "Have you considered using the Internet to research those questions?" Responses were transcribed and subjected to qualitative thematic analysis. Two researchers independently reviewed the data and performed inductive coding to identify recurring patterns and themes. Through consensus and iterative discussions, responses were systematically organized into major concern areas. Our analysis revealed that 60% of participants were unaware of any patient education materials relating to pregnancy, and 80% reported relying on the Internet as

their primary source of information. One participant highlighted, *"I searched information on the health of the mother and baby, the growth of the baby, and the diet that the mother has to follow. First-time mothers, in particular, are often thoroughly confused about whether it is okay to eat certain foods or sleep in specific positions."* The most common themes identified through this process included postpartum care, pregnancy health and complications, prenatal preparation and support, diet and nutrition, mental and emotional wellbeing, birth preferences and experiences, and practical preparations for baby.

Data collection and thematic analysis were conducted iteratively, with two researchers independently reviewing interview responses and identifying emergent themes. Recruitment and interviews continued until thematic saturation was reached, defined as the point at which no new major themes or substantive concerns emerged from additional participant input. This process ensured that the thematic domains generated fully reflected the range and diversity of participant experiences and concerns.

3.3 Formulation and validation of representative questions

Based on these identified themes, we formulated a set of 50 representative questions, designed to reflect the full breadth and diversity of concerns voiced by participants. This preliminary question set was then reviewed and validated by a panel of experienced healthcare professionals (doctors), who evaluated each question for relevance, clarity, and alignment with the underlying themes. Feedback from the doctors was incorporated to refine and finalize the question set. The resulting 50 questions thus represent a balanced and validated sample of commonly encountered pregnancy-related concerns, which were subsequently used for evaluating chatbot performance with both participant groups.

3.4 Adaptation of NHS pregnancy book content for chatbot responses

For the development of the chatbot's knowledge base and responses, we adopted the entire content from the NHS guidebook, *The Pregnancy Book: Your Complete Guide to a Healthy Pregnancy, Labour and Childbirth, and the First Weeks with Your New Baby*. This comprehensive, evidence-based guide was selected for its breadth, clinical reliability, and national standard status in patient education. The full text was integrated into the chatbot's responses, ensuring that users had access to holistic and authoritative information on pregnancy-related topics.

During adaptation for chatbot use, language from the original guide was simplified where necessary to accommodate the reading levels of our target population, specifically individuals with less than a high school education. Simplification involved paraphrasing technical terms and complex sentences to enhance understanding, while care was taken to preserve the accuracy and intent of the medical advice. All adapted content underwent review by healthcare professionals to ensure fidelity to the original guidance and suitability for lay users before deployment within the chatbot platform.

3.5 Operationalization and assessment of comprehensibility–accuracy–readability (CAR) scores

We adopted the evaluation process detailed in a recent study on developing AI-generated medical responses for cancer patients (Lee et al., 2024), which assessed responses based on the comprehensibility–accuracy–and readability (CAR) scores.

Comprehensibility was defined as the degree to which a response could be easily understood by a lay audience, emphasizing logical flow, coherence, and absence of ambiguity; it was operationalized using PEMAT Understandability criteria (Shoemaker et al., 2014). Accuracy referred to the factual and clinical correctness of the information, its alignment with current obstetric guidelines, and its relevance to the question; it was operationalized by comparing each response against authoritative sources (e.g., the NHS guidelines) and assigning a rating for correctness and completeness (Shepperd and Charnock, 2002). Readability reflected whether language, vocabulary, and sentence structure were suitable for individuals with less than a high school education, avoiding technical jargon and unnecessary complexity; it was operationalized via Flesch–Kincaid Grade Level, targeting a score $\leq$ 8th-grade level for all responses (Kruse et al., 2020). All responses were rated independently by evaluators using a standardized 4-point Likert scale: 1 = Insufficient, 2 = Moderate, 3 = Good, 4 = Very Good.

Readability of chatbot responses was evaluated using both quantitative and user-centered approaches. For each response, we checked sentence length and structure to ensure the text was simple, concise, and suitable for non-specialist readers. In addition, participants were asked to rate the ease of reading and list any words or phrases they found difficult or unfamiliar. This method follows established patient information evaluation frameworks, which recommend combining simple linguistic analysis with direct feedback from users to improve accessibility and identify barriers to understanding (Garner et al., 2011).

To control for variations in interpretation and engagement, all participants evaluated chatbot responses to the same set of pre-defined questions, rather than chatting freely. Before the evaluation, all evaluators, both doctors and women, were taught how to use the CAR framework. This training included going over sample chatbot answers together and discussing how each of the three CAR categories should be scored. The goal was to make sure everyone understood the criteria in the same way. After initial training, evaluators scored some example responses. If they gave different scores for the same example, the team discussed why. Those discussions helped clarify any ambiguities and align everyone's judgments. When evaluators scored the real chatbot responses, each answer was stripped of any information identifying who authored it or under what circumstances it was generated. This blinding prevents any conscious or unconscious bias from affecting their scores.

3.6 Statistical analysis and visualization

Following data collection, we conducted Mann–Whitney *U* tests (Ruxton, 2006) to compare the responses of doctors and women on CAR across four content areas: (1) postpartum care, (2) pregnancy health and complications, (3) diet and nutrition, and (4) mental and

emotional wellbeing. This non-parametric test was chosen because of its suitability for ordinal data and its robustness when dealing with non-normal distributions or unequal group sizes.

Between-group comparisons for each rating domain (comprehensibility, accuracy, and readability) were performed using the Mann–Whitney *U* test, appropriate for ordinal 4-point Likert ratings and non-normal distributions. For each comparison, we report the standardized effect size $r = |Z|/\sqrt{N}$ ($N = n1 + n2$) (Ialongo, 2016), and to complement r, we additionally report Cliff's delta (Δ) to express dominance (stochastic superiority) between groups (Rahlfs and Zimmermann, 2019). Exact/asymptotic *p*-values and tie handling followed the software defaults, and Bonferroni correction was applied across the four domains per outcome. Inter-rater reliability indices were not computed for participant ratings, as respondents functioned as independent raters providing their perceptions, rather than interchangeable judges of the same items for agreement. Thus, our analytic focus was on comparing group rating distributions rather than assessing inter-rater reliability (Figure 1).

By leveraging Google Colab, we combined non-parametric statistical analysis with radar charts to graphically present our findings (Figure 2). Radar charts overlay multiple quantitative dimensions in a single, coherent shape, enabling immediate visual detection of systematic differences between groups and across variables. This multidimensional "polygon" format supports holistic comparison without requiring several separate graphs (Schlee et al., 2017). By mapping each content area axis, postpartum care, pregnancy health and complications, diet and nutrition, mental and emotional wellbeing, and plotting mean-rank CAR scores for doctors and women as overlaid polygons, radar plots reveal each group's distinct strengths and weaknesses in a single view. The relative compactness or elongation of polygons across axes visually encodes dimensional uniformity versus variability (Mason et al., 2024). Radar plots emphasize axes where group polygons diverge most, directing readers to specific content areas that may need targeted improvement (for instance, areas where doctors rate comprehensibility lower than women). This intuitive, at-a-glance interpretation enhances readability for both methodological and clinical audiences. To facilitate interpretation of group comparisons across multiple response domains (comprehensibility, accuracy, and readability), radar charts were constructed for each study group (doctors and women). For each chart, the mean CAR scores for postpartum care, pregnancy health and complications, diet and nutrition, and mental and emotional wellbeing were plotted on separate axes radiating from a common center. This allowed simultaneous visualization of performance across all domains for each group, highlighting strengths and weaknesses in chatbot responses. Radar charts support intuitive comparisons by visually expressing where profiles overlap, diverge, or show relative advantage, thus complementing the statistical tables.

3.7 Informed consent

Informed consent was obtained from all participants, both women and healthcare professionals, before their involvement in the study. Participants received detailed information about the study's purpose, procedures, potential risks, and benefits. Participation was entirely voluntary, and individuals could withdraw at any time without

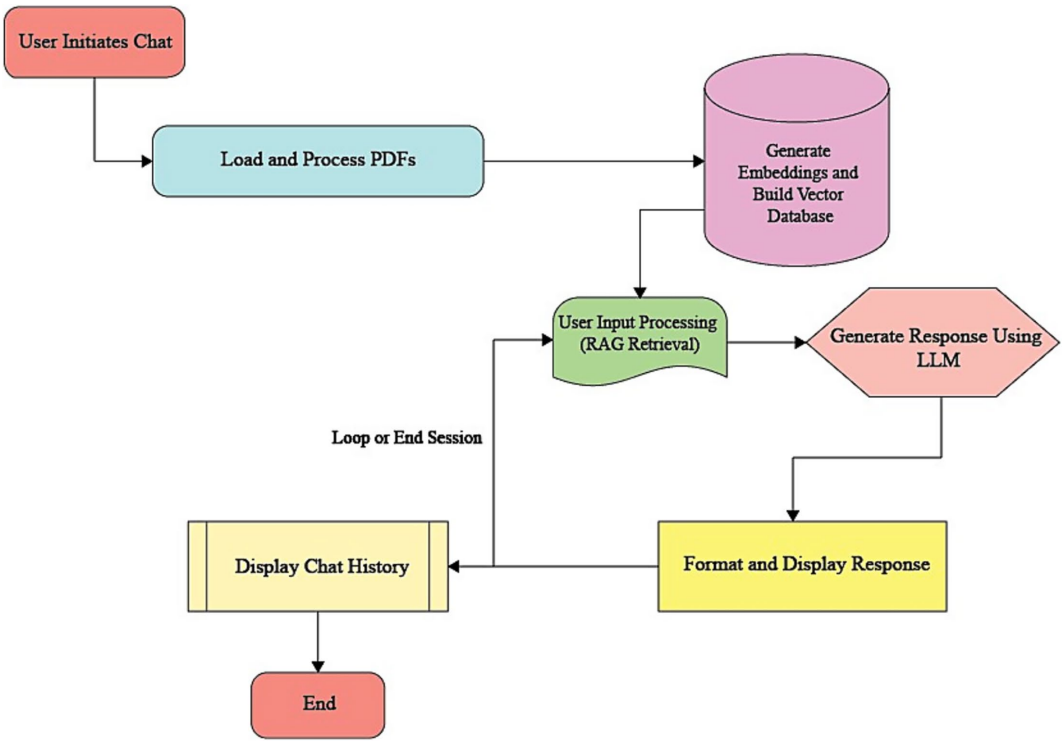

FIGURE 1
DIAN app architecture.

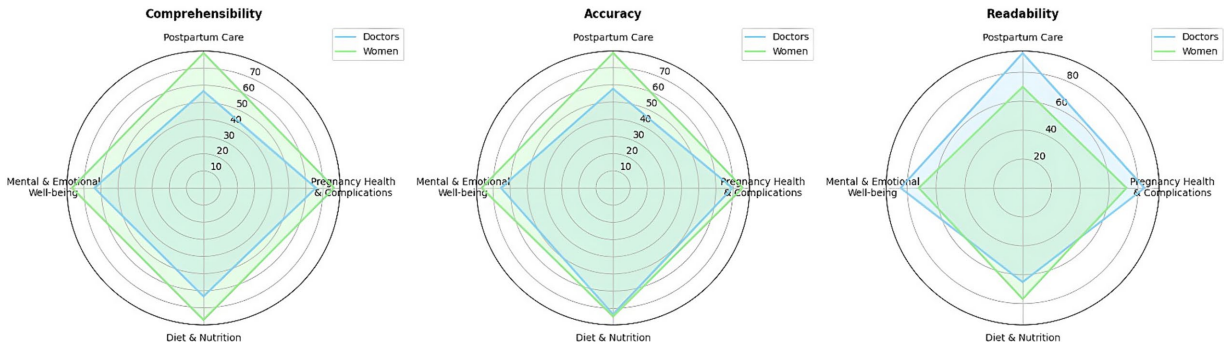

FIGURE 2
Radar chart comparing doctors' and women's ratings.

consequence. Confidentiality and anonymity were assured for all responses and data collected.

4 Results

This research consisted of two study groups comprising 119 women and 29 doctors. The female study group majority aged 18–25 years (52.1%), followed by 26–40 years (42.0%) and 41–55 years (5.9%). The majority of patient participants had completed school (53.8%), with additional representation from undergraduate (29.4%) and postgraduate (16.8%) educational levels. The doctor group consisted of 21 women and 8 men, aged 26–40 years (65.5%) or 41–55 years (34.5%). The sample sizes (n = 29 clinicians; n = 119 women) reflect recruitment feasibility during the study window. *Post-hoc* power estimates based on observed effects indicated achieved power of ~0.48–0.57 for the largest contrasts (postpartum care), suggesting the study may be underpowered; we therefore advise larger, balanced samples in future studies. The majority of doctors held postgraduate degrees or higher qualifications. All participants were recruited from a region where English is not the primary language, representing predominantly non-native English speakers. Participant characteristics are detailed in Table 3.

Doctors' and women's ratings on three important dimensions, readability, accuracy, and comprehension across four content areas (1) postpartum care, (2) pregnancy health and complications, (3) diet and nutrition, and (4) mental and emotional wellbeing were compared using a series of Mann–Whitney U tests. Table 4 (comprehensibility), Table 5 (accuracy), and Table 6 (readability) display the findings.

According to Table 4, the only area where there was a statistically significant difference in the comprehension of women and doctors was postpartum care (U = 1206.50, Z = −2.524, p = 0.012). Compared to doctors (mean rank = 56.60), doctors found postpartum care information easier to understand (mean rank = 78.86). To supplement the statistical significance, the effect size for comprehensibility in postpartum care was r = 0.207 (small to moderate) and Cliff's delta = 0.301, indicating that the practical difference between women and doctors was small to moderate. Interestingly, no significant differences were found for the other three content areas: mental and emotional wellbeing, diet and nutrition, and pregnancy health and complications. This suggests that both groups generally had similar opinions regarding the simplicity or ease of understanding the content in these areas.

Since comprehensibility gauges how easily information can be understood, women's higher scores on postpartum care may reflect their familiarity and experience with the topic, while physicians may evaluate comprehensibility in comparison with a clinical standard. This distinction emphasizes the importance of conveying postpartum care information in a manner that is both clinically accurate and understandable to all audiences, including non-professionals.

The accuracy results (Table 5) also show a significant difference only for postpartum care (U = 1239.00, Z = −2.370, p = 0.018), where women had higher mean ranks (78.59) than physicians (57.72). The effect size for accuracy in postpartum care was r = 0.195 (small) and Cliff's delta = 0.282, suggesting the practical difference was small. Diet and nutrition, mental and emotional wellbeing, and pregnancy health and complications did not show any statistically significant differences.

Accuracy refers to whether the content is factually correct, reliable, and in line with current medical or experiential knowledge. Because of their close connection to and involvement in postpartum events, women may perceive postpartum treatment to be more accurate. In contrast, physicians who use a more rigorous clinical lens might examine the same material more closely. These results suggest a potential discrepancy between the evaluation of factual correctness by professional and lay audiences, underscoring the significance of sophisticated evidence-based communication techniques in postpartum care.

According to Table 6, there is one more significant difference in reading between women and doctors in postpartum care (U = 1181.50, Z = −2.639, p = 0.008). In contrast to women (mean rank = 69.93), doctors notably thought the postpartum care language was easier to understand (mean rank = 93.26). No discernible variations in readability scores were observed across other subject areas. The corresponding effect size for readability in postpartum care was r = 0.217 (small to moderate) and Cliff's delta = 0.315, indicating the practical magnitude of differences was small to moderate.

Readability measures the ease of reading and processing a text. The use of clinical terminology or structure that is more in line with a medical professional's reading expectations may be the reason for the higher ratings given by physicians for postpartum care. In contrast, women, who might choose simpler, informal language, scored worse on reading tests. This disparity highlights the importance of adjusting presentation style and text complexity to the target audience, especially when discussing delicate subjects such as postpartum care.

TABLE 3 Demographic characteristics of study participants.

Study group	Gender (F/M)	Age 18–25	Age 26–40	Age 41–55	School	UG	PG	PhD/MD	*N*
Women	119/0	62	50	7	64	35	20	0	119
Doctors	21/8	0	19	10	–	–	18	1	29

TABLE 4 Comprehensibility.

Group questions (factors)	Mean rank		Mann–Whitney *U* test	*Z*-value	*P*-value	p_Bonf	*r*	Effect size magnitude	Cliff's Δ	Direction
	Doctor	Women								
Postpartum care	56.60	78.86	1206.500	−2.524	0.012	0.144	0.207	Small–moderate	0.301	Patients > doctors
Pregnancy health and complications	65.86	76.61	1475.000	−1.219	0.223	1	0.1	Small	0.145	Patients > doctors
Diet and nutrition	63.33	77.08	1401.500	−1.575	0.115	1	0.129	Small	0.188	Patients > doctors
Mental and emotional wellbeing	63.93	77.08	1419.000	−1.485	0.138	1	0.122	Small	0.178	Patients > doctors

TABLE 5 Accuracy.

Group questions (factors)	Mean rank		Mann–Whitney *U* test	*Z*-value	*P*-value	p_Bonf	*r*	Effect size magnitude	Cliff's Δ	Direction
	Doctor	Women								
Postpartum care	57.72	78.59	1239.000	−2.370	0.018	0.216	0.195	Small	0.282	Patients > doctors
Pregnancy health and complications	69.66	75.68	1585.000	−0.680	0.497	1	0.056	Negligible	0.081	Patients > doctors
Diet and nutrition	73.50	74.74	1696.500	−0.141	0.888	1	0.012	Negligible	0.017	Patients > doctors
Mental and emotional wellbeing	65.41	76.71	1462.000	−1.274	0.203	1	0.105	Small	0.153	Patients > doctors

TABLE 6 Readability.

Group questions (factors)	Mean rank		Mann–Whitney *U* test	*Z*-value	*P*-value	p_Bonf	*r*	Effect size magnitude	Cliff's Δ	Direction
	Doctor	Women								
Postpartum care	93.26	69.93	1181.500	−2.639	0.008	0.096	0.217	Small–moderate	0.315	Doctors > patients
Pregnancy health and complications	84.53	72.05	1434.500	−1.416	0.157	1	0.116	Small	0.169	Doctors > patients
Diet and nutrition	65.09	76.79	1452.500	−1.338	0.181	1	0.11	Small	0.158	Patients > doctors
Mental and emotional wellbeing	84.60	72.04	1432.500	−1.422	0.155	1	0.117	Small	0.17	Doctors > patients

Across all assessed domains, reported effect sizes were predominantly small, indicating that while some group differences reached statistical significance, their magnitude was limited in practical terms. This underscores the value of reporting effect sizes alongside *p*-values.

Radar charts provided an immediate, holistic visualization of performance differences between doctors and women across all evaluated domains: comprehensibility, accuracy, and readability. By mapping mean scores for each content area (postpartum care, pregnancy health and complications, diet and nutrition, and mental and emotional wellbeing) onto separate axes and overlaying group polygons, the charts allowed simultaneous comparison of group profiles in a single, intuitive view.

This format made dimensional strengths and weaknesses visibly apparent: For example, the marked expansion of the women's polygon on the postpartum care axis for comprehensibility and accuracy highlighted their higher ratings in this domain, while the pronounced extension of the doctors' polygon in readability for the same axis showcased their relative advantage there. The visual contrast between polygons directly echoed patterns detected in the statistical analysis, emphasizing where divergences were most substantial and supporting quick identification of domains needing further improvement.

5 Discussion

We examined the performance of the DIAN chatbot in advising women. As we implemented the grading technique based on prior studies, our chatbot promises to be the best integration method while increasing patient education. Postpartum care was the only area where the ratings of doctors and women varied significantly across all three criteria. Women rated postpartum care information higher for accuracy and comprehensibility, indicating that content tailored to postpartum experiences may be better suited to their unique needs and opinions. Postpartum care is not just informational but deeply emotional, with women reporting explicit desire for more emotionally attuned, patient-centered, and understandable support (Roberts et al., 2025). In our study, divergence in postpartum care response within CAR metrics highlights that deeply emotional and precisely tailored communication can have a significant impact. Postpartum-related issues present unique clinical complexities (Fox et al., 2018). Both women and clinicians have distinct expectations and needs in postpartum dialogs, and harder to standardize, especially for AI chatbots. Postpartum content often addresses issues of trauma, depression, and abuse, making it not only technically complex but also deeply emotionally sensitive (Islam et al., 2020). Even small failures in clarity, trust, or appropriateness are amplified because postpartum mothers are acutely attuned to the tone, nuance, and completeness of medical advice (Amar and Sejfović, 2023). The emotionally charged, multidimensional nature of postpartum care, requiring both technical information and emotional reassurance, logically leads to gaps in CAR metrics, with women participants accepting simple content while doctors expect clinical information from the same postpartum material.

Particularly for first-time mothers, the uneven quality or questionable sources of Internet information regarding pregnancy and Childcare can be frightening (Chua et al., 2023). Our findings indicate that postpartum care is a crucial topic with notable variations in readability, correctness, and comprehensibility between expectant women and physicians. These variations imply that different audiences may have different perceptions or interpretations of postpartum content. A previous study with Rosie Chatbot (Nguyen et al., 2024) emphasizes the value of accurate and culturally sensitive postpartum information by using approved, on-demand content. Our research supports the basic idea that Rosie's design is a useful digital tool for treating postpartum depression. Our work underlines the value of user-centered design and iterative feedback in creating a text-based conversational agent (Calvo et al., 2023). This approach aligns with our method of determining whether various end users find the chatbot content comprehensible and trustworthy. Together, these findings reinforce the importance of designing conversational agents that adapt dynamically to user feedback, employ brief and comprehensible messaging, and remain tightly aligned with clinical best practices.

The findings from our chatbot evaluations spanning postpartum care, pregnancy health and complications, diet and nutrition, and mental and emotional wellbeing mirror key themes from the recent study by Kaphingst et al. (2024), which demonstrated that automated conversational agents can be equivalent to standard of care (SOC) approaches in delivering key health information. In their study, a chatbot guided patients through cancer genetic services, achieving completion rates for pretest counseling and test uptake like those seen with in-person appointments. This equivalence is highly relevant to our context. Although their focus was on cancer risk assessment, the takeaway is that a well-designed chatbot can successfully convey complex medical content, such as postpartum guidelines or nutrition education, to a broad patient audience.

The review of dementia-focused chatbots highlighted limitations in achieving natural, adaptive dialog and comprehensive content delivery (Ruggiano et al., 2021). Their findings imply that many existing systems struggle to balance technical accuracy and user-friendly communication. Our results reinforce this notion that optimizing comprehensibility and accuracy for lay audiences while maintaining readability for expert users is crucial. Both studies underscore that successful chatbot interventions must harmonize evidence-based, accessible content with adaptable conversation flows to meet users' diverse needs.

Our chatbot study and DR-COVID (Yang et al., 2023) work underscores the promise of AI-driven conversational agents in health education. Our study demonstrates that chatbot responses achieve high comprehensibility and accuracy for lay users, whereas clinicians value technical and scientific readability. Similarly, DR-COVID's ensemble NLP approach achieved robust overall accuracy (0.838) and top-3 accuracy (0.922) in delivering COVID-19 information across multiple languages. Despite their differing domains, both studies highlight that adaptive, evidence-based chatbot systems can effectively translate complex medical information into user-friendly, reliable guidance.

Our findings highlight the importance of modifying health information for various audiences. For women in the lay audience, accuracy and comprehensibility can be improved by using clear language and practical relevance. Clinical Audience (Doctors) Professional terminology or structure can enhance reading. There were no notable differences between the two groups in terms of diet and nutrition, mental and emotional wellbeing, or pregnancy health and complications. These results imply that there might be a more general agreement or useful resources available for these subjects outside of postpartum care. Ongoing improvement, however, might make use of methods such as audience segmentation, natural language

processing, or dynamic text production to better tailor information delivery to the unique requirements of each subgroup in future studies.

6 Conclusion

Chatbots are not human clinicians (Manole et al., 2024). They cannot fully replicate a human provider's ability to interpret complex clinical situations or personalize advice beyond the knowledge base they are trained on (Frodl et al., 2024). Because chatbots rely on existing data sources (Chow et al., 2025), inaccuracies or biases (Van Poucke, 2024) can be perpetuated unless those sources are carefully vetted. Using patient education content, this study adds to the expanding corpus of information on chatbot technology. Research must continue as chatbot technology develops quickly to stay updated with new trends, user behavior, and societal ramifications (Følstad et al., 2021). Chatbots offer a quick, easy, and varied way to exchange information (Yu et al., 2023). It is crucial to employ chatbots appropriately, solve ethical issues, and carry out additional research to fully reap the benefits of this technology in medical and health sciences (Razak et al., 2023).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The study was conducted in accordance with applicable local legislation and institutional requirements. All participants provided written informed consent prior to participation. No Sensitive data collected from the participants.

Author contributions

PVa: Writing – original draft, Formal analysis, Methodology, Data curation, Investigation, Writing – review & editing, Software. PVe: Methodology, Supervision, Conceptualization, Software, Investigation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is funded by the Vellore Institute of Technology, Vellore, India.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. The authors acknowledge the use of AI-assisted technologies in the preparation of this manuscript to enhance readability, improve language clarity, and assist with grammar checking and sentence restructuring. The final manuscript was reviewed and approved by the authors, who take full responsibility for its content.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abdullah, Y., Alokozai, A., Mathew, A. J., Stamm, M. A., and Mulcahey, M. K. (2022). Patient education materials found via Google search for shoulder arthroscopy are written at too-high of a Reading level. *Arthrosc. Sports Med. Rehabil.* 4, e1575–e1579. doi: 10.1016/j.asmr.2022.04.034

Agrawal1, A. M., Shinde2, R. P., Bhukya3, V. K., Chakraborty4, A., Shah5, B., Shukla6, T., et al. (2025). Conversation AI dialog for Medicare powered by finetuning and retrieval augmented generation. *ResMilitaris, Social Science Journal*, 12, 2265–6294. doi: 10.48550/arXiv.2502.02249

Akpan, I. J., Kobara, Y. M., Owolabi, J., Akpan, A. A., and Offodile, O. F. (2025). Conversational and generative artificial intelligence and human–chatbot interaction in education and research. *Int. Trans. Oper. Res.* ResMilitaris, Social Science Journal, 32, 1251–1281. doi: 10.1111/ITOR.13522

Amar, A. V., and Sejfović, H. (2023). Perceived social support, newborn temperament and socioeconomic status in postpartum depression: report from Southwest Serbia. *Arch. Psychiatry Psychother.* 25, 33–41. doi: 10.12740/APP/152779

Elbanna, A. R. (2007). Implementing an integrated system in a socially dis-integrated enterprise: A critical view of ERP enabled integration, *Information Technology.* 20:121–139. doi: 10.1108/09593840710758040

Bavaresco, R., Silveira, D., Reis, E., Barbosa, J., Righi, R., Costa, C., et al. (2020). Conversational agents in business: a systematic literature review and future research directions. *Comput Sci Rev* 36:100239. doi: 10.1016/j.cosrev.2020.100239

Calvo, R. A., Peters, D., Moradbakhti, L., Cook, D., Rizos, G., Schuller, B., et al. (2023). Assessing the feasibility of a text-based conversational agent for asthma support: protocol for a mixed methods observational study. *JMIR Res. Protocol* 12:2965. doi: 10.2196/42965

Chow, R., Suen, K. Y., and Lam, A. Y. S. (2025). On leveraging large language models for multilingual intent discovery. *ACM Trans. Manag. Inf. Syst.* 16:17. doi: 10.1145/3688400

Chu, S. N., and Goodell, A. J. (2024) Synthetic patients: simulating difficult conversations with multimodal generative AI for medical education. ArXiv under Cornell University stewardship, New York City . doi: 10.48550/arXiv.2405.19941

Chua, J. Y. X., Choolani, M., Chee, C. Y. I., Chan, Y. H., Lalor, J. G., Chong, Y. S., et al. (2023). Insights of parents and parents-to-be in using chatbots to improve their preconception, pregnancy, and postpartum health: a mixed studies review. *J. Midwifery Womens Health* 68, 480–489. doi: 10.1111/jmwh.13472

Chung, K., Cho, H. Y., and Park, J. Y. (2021). A Chatbot for perinatal womens and partners⇔ obstetric and mental health care: development and usability evaluation study. *JMIR Med. Inform.* 9:e18607. doi: 10.2196/18607

Du, Y., Abbas, S., Liu, X., Wang, X., He, X., Wei, J., et al. (2019) Application of artificial intelligence to the public health education, *Frontiers in Public Health. Lausanne:*, Switzerland.

Følstad, A., Araujo, T., Law, E. L. C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., et al. (2021). Future directions for Chatbot research: an interdisciplinary research agenda. *Computing* 103, 2915–2942. doi: 10.1007/s00607-021-01016-7

Fox, M., Sandman, C. A., Davis, E. P., and Glynn, L. M. (2018). A longitudinal study of women's depression symptom profiles during and after the postpartum phase. *Depress. Anxiety* 35, 292–304. doi: 10.1002/DA.22719

Frodl, A., Fuchs, A., Yilmaz, T., Izadpanah, K., Schmal, H., and Siegel, M. (2024). ChatGPT as a source for patient information on patellofemoral surgery—a comparative study amongst laymen, doctors, and experts. *Clin Pract* 14, 2376–2384. doi: 10.3390/clinpract14060186

Fotheringham, D., and Wiles, M. A. (2023). The effect of implementing chatbot customer service on stock returns: An event study analysis. *Journal of the Academy of Marketing Science*, 51, 802–822. doi: 10.1007/s11747-022-00841-2

Ganesh, S., Grandhi, A. S. K., Konduri, P., Samudrala, P. K., and Nemmani, K. V. S. (2022). Advancing health care via artificial intelligence: from concept to clinic. *Eur. J. Pharmacol.* 934. doi: 10.1016/j.ejphar.2022.175320

Gargari, O. K., and Habibi, G. (2025). Enhancing medical AI with retrieval-augmented generation: a Mini narrative review. *Digital Health* 11:20552076251337176. doi: 10.1177/20552076251337177

Garner, M., Ning, Z., and Francis, J. (2011). A framework for the evaluation of patient information leaflets. *Health Expect.* 15, 283–294. doi: 10.1111/J.1369-7625.2011.00665.X

Gkinko, L., and Elbanna, A. (2023). The appropriation of conversational AI in the workplace: a taxonomy of AI chatbot users. *Int. J. Inf. Manag.* 69:2568. doi: 10.1016/j.ijinfomgt.2022.102568

Hong, G., Smith, M., and Lin, S. (2022). The AI will see You now: feasibility and acceptability of a conversational AI medical interviewing system. *JMIR Formative Res.* 6:e37028. doi: 10.2196/37028

Ialongo, C. (2016). Understanding the effect size and its measures. *Biochem. Med.* 26, 150–163. doi: 10.11613/BM.2016.015

Islam, M. J., Broidy, L., Mazerolle, P., Baird, K., Mazumder, N., and Zobair, K. M. (2020). Do maternal depression and self-esteem moderate and mediate the association between intimate partner violence after childbirth and postpartum suicidal ideation? *Arch. Suicide Res.* 24, 609–632. doi: 10.1080/13811118.2019.1655507

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., et al. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* 2, 230–243. doi: 10.1136/svn-2017-000101

Kaphingst, K. A., Kohlmann, W. K., Chambers, R. L., Bather, J. R., Goodman, M. S., Bradshaw, R. L., et al. (2024). Uptake of cancer genetic services for chatbot vs. standard-of-care delivery models: the BRIDGE randomized clinical trial. *JAMA Netw. Open* 7:e2432143. doi: 10.1001/jamanetworkopen.2024.32143

Kruse, J., Toledo, P., Belton, T. B., Testani, E. J., Evans, C. T., Grobman, W. A., et al. (2020). Readability, content, and quality of COVID-19 patient education materials from academic medical centers in the United States. *Am. J. Infect. Control* 49:690. doi: 10.1016/J.AJIC.2020.11.023

Lee, J. w., Yoo, I. S., Kim, J. H., Kim, W. T., Jeon, H. J., Yoo, H. S., et al. (2024). Development of AI-generated medical responses using the chatGPT for cancer patients. *Comput. Methods Prog. Biomed.* 254:108302. doi: 10.1016/J.CMPB.2024.108302

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in neural information processing systems 2020-December. Available online at: https://arxiv.org/pdf/2005.11401 (accessed December 30, 2024).

Maeda, E., Miyata, A., Boivin, J., Nomura, K., Kumazawa, Y., Shirasawa, H., et al. (2020). Promoting fertility awareness and preconception health using a Chatbot: a randomized controlled trial. *Reprod. Biomed. Online* 41, 1133–1143. doi: 10.1016/j.rbmo.2020.09.006

Manole, A., Cârciumaru, R., Brînzaș, R., and Manole, F. (2024). Harnessing AI in anxiety management: a Chatbot-based intervention for personalized mental health support. *Information* 15:768. doi: 10.3390/info15120768

Mason, L., Otero, M., and Andrews, A. (2024). Analyzing the functional interdependence of verbal behavior with multiaxial radar charts. *Perspect. Behav. Sci.* 47, 471–498. doi: 10.1007/S40614-024-00404-6

Mondal, H., Panigrahi, M., Mishra, B., Behera, J. K., and Mondal, S. (2023). A pilot study on the capability of artificial intelligence in preparation of patients' educational materials for Indian public health issues. *J. Family Med. Prim. Care* 12, 1659–1662. doi: 10.4103/JFMPC.JFMPC_262_23

Nasra, M., Jaffri, R., Pavlin-Premrl, D., Kok, H. K., Khabaza, A., Barras, C., et al. (2025). Can artificial intelligence improve patient educational material readability? A systematic review and narrative synthesis. *Intern. Med. J.* 55, 20–34. doi: 10.1111/IMJ.16607;WGROUP:STRING:PUBLICATION

Nattam, A., Vithala, T., Wu, T. C., Bindhu, S., Bond, G., Liu, H., et al. (2023). Assessing the readability of online patient education materials in obstetrics and gynecology using traditional measures: comparative analysis and limitations. *J. Med. Internet Res.* 25:e46346. doi: 10.2196/46346

Nguyen, Q. C., Aparicio, E. M., Jasczynski, M., Doig, A. C., Yue, X., Mane, H., et al. (2024). Rosie, a health education question-and-answer chatbot for new mothers: randomized pilot study. *JMIR Form. Res.* 8. doi: 10.2196/51361

Oche, A. J., Folashade, A. G., Ghosal, T., and Biswas, A.. (2025). "A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions.".

Rahlfs, V., and Zimmermann, H. (2019). Effect size measures and their benchmark values for quantifying benefit or risk of medicinal products. *Biometric. J.* 61, 973–982. doi: 10.1002/BIMJ.201800107

Razak, N. I. A., Yusoff, M. F. M., and Rahmat, R. W. O. K. (2023). ChatGPT review: a sophisticated chatbot models in medical and health-related teaching and learning. *Malays. J. Med. Health Sci.* 19, 98–108. doi: 10.47836/mjmhs.19.s12.12

Roberts, T. P., Nowakowski, E. E., Troyan, T. H., Kroh, S. J., Wanaselja, A. M., Gopalan, P. R., et al. (2025). Improving psychological and social support needs after traumatic birth: a qualitative study. *J. Affect. Disord. Rep.* 19:100849. doi: 10.1016/J.JADR.2024.100849

Rokicki, S., Gobburu, A., Weidner, M., Azam, N., Jansen, M., Rivera-Núñez, Z., et al. (2025). Barriers and strategies for recruitment of pregnant women in contemporary longitudinal birth cohort studies. *BMC Med. Res. Methodol.* 25, 1–10. doi: 10.1186/S12874-025-02570-W/TABLES/3

Ruggiano, N., Brown, E. L., Roberts, L., Suarez, C. V. F., Luo, Y., Hao, Z., et al. (2021). Chatbots to support people with dementia and their caregivers: systematic review of functions and quality. *J. Med. Internet Res.* 23. doi: 10.2196/25006

Ruxton, G. D. (2006). The unequal variance T-test is an underused alternative to student's t-test and the Mann-Whitney U test. *Behav. Ecol.* 17, 688–690. doi: 10.1093/beheco/ark016

Sagstad, M. H., Morken, N. H., Lund, A., Dingsør, L. J., Nilsen, A. B. V., and Sorbye, L. M. (2022). Quantitative user data from a Chatbot developed for women with gestational diabetes mellitus: observational study. *JMIR Format. Res.* 6:e28091. doi: 10.2196/28091

Saunders, C. H., Petersen, C. L., Durand, M.-A., Bagley, P. J., and Elwyn, G. (2018). Bring on the machines: could machine learning improve the quality of patient education materials? A systematic search and rapid review. *JCO Clin. Cancer Informat.* 2, 1–16. doi: 10.1200/CCI.18.00010

Schlee, W., Hall, D. A., Edvall, N. K., Langguth, B., Canlon, B., and Cederroth, C. R. (2017). Visualization of global disease burden for the optimization of patient management and treatment. *Front. Med.* 4:86. doi: 10.3389/FMED.2017.00086/FULL

Schwalbe, N., and Wahl, B. (2020). Artificial intelligence and the future of Global Health. *Lancet* 395, 1579–1586. doi: 10.1016/S0140-6736(20)30226-9

Shepperd, S., and Charnock, D. (2002). DISCERN: why DISCERN? *Health Expect.* 1, 134–135. doi: 10.1046/J.1369-6513.1998.0112A.X

Shi, W., Zhuang, Y., Zhu, Y., Iwinski, H., Wattenbarger, M., and Wang, M. D.. (2023). Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. ACM-BCB 2023 – 14th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics.

Shin, M., Song, J., Kim, M. G., Yu, H. W., Choe, E. K., and Chai, Y. J. (2025). Thyro-GenAI: a chatbot using retrieval-augmented generative models for personalized thyroid disease management. *J. Clin. Med.* 14:2450. doi: 10.3390/JCM14072450/S1

Slatore, C. G., Kulkarni, H. S., Corn, J., and Sockrider, M. (2016). Improving health literacy, Health Literacy Research and Practice, Thorofare, USA. doi: 10.3928/24748307-20220420-01

Shoemaker, S. J., Wolf, M. S., and Brach, C. (2014). Development of the patient education materials assessment tool (PEMAT): a new measure of understandability and Actionability for print and audiovisual patient information. *Patient Educ. Couns.* 96, 395–403. doi: 10.1016/J.PEC.2014.05.027

Speirs, K. E., Grutzmacher, S. K., Munger, A. L., and Messina, L. A. (2015). Recruitment and retention in an SMS-based health education program: Lessons learned from Text2BHealthy. *Health Informatics Journal*, 22, 651–658 doi: 10.1177/1460458215577995

Tian, F., Ganguly, D., and Macdonald, C. (2025). Is relevance propagated from retriever to generator in RAG? *Lect. Notes Comput. Sci* 15572, 32–48. doi: 10.1007/978-3-031-88708-6_3

Van Poucke, M. (2024). ChatGPT, the perfect virtual teaching assistant? Ideological Bias in learner-Chatbot interactions. *Comput. Compos.* 73:102871. doi: 10.1016/j.compcom.2024.102871

Yang, L. W. Y., Ng, W. Y., Lei, X., Tan, S. C. Y., Wang, Z., Yan, M., et al. (2023). Development and testing of a multi-lingual natural language processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: a multi-center study. *Front. Public Health* 11:3466. doi: 10.3389/fpubh.2023.1063466

Yu, C. S., Hsu, M. H., Wang, Y. C., and You, Y. J. (2023). Designing a Chatbot for helping parenting practice. *Appl. Sci.* 13:1793. doi: 10.3390/app13031793