scientific reports

Check for updates

OPEN

Quality assessment of large language models' output in maternal health

Henrique A. Lima[3], Pedro H. F. S. Trocoli-Couto[3], Zorays Moazzam[7], Leonardo C. D. Rocha[5], Adriana Pagano[4], Felipe F. Martins[6], Lucas T. Brabo[6], Zilma S. N. Reis[3], Lisa Keder[2], Aliya Begum[8], Marcelo H. Mamede[3], Timothy M. Pawlik[1] & Vivian Resende[3]✉

Optimising healthcare is linked to broadening access to health literacy in Low- and Middle-Income Countries. The safe and responsible deployment of Large Language Models (LLMs) may provide accurate, reliable, and culturally relevant healthcare information. We aimed to assess the quality of outputs generated by LLMs addressing maternal health. We employed GPT-4, GPT-3.5, GPT-3.5 custom, Meditron-70b. Using mixed-methods, cross-sectional survey approach, specialists from Brazil, United States, and Pakistan assessed LLM-generated responses in their native languages to a set of three questions relating to maternal health. Evaluators assessed the answers in technical and non-technical scenarios. The LLMs' responses were evaluated regarding information quality, clarity, readability and adequacy. Of the 47 respondents, 85% were female, mean age of 50 years old, with a mean of 19 years of experience (volume of 110 assisted pregnancies monthly). Scores attributed to answers by GPT-3.5 and GPT-4 were consistently higher [Overall, GPT-3.5, 3.9 (3.8–4.1); GPT-4.0, 3.9 (3.8–4.1); Custom GPT-3.5, 2.7 (2.5–2.8); Meditron-70b, 3.5 (3.3–3.6); $p = 0.000$]. The responses garnered high scores for clarity (Q&A-1 3.5, Q&A-2 3.7, Q&A-3 3.8) and for quality of content (Q&A-1 3.2, Q&A-2 3.2, Q&A-3 3.7); however, they differed by language. The commonest limitation to quality was incomplete content. Readability analysis indicated that responses may require high educational level for comprehension. Gender bias was detected, as models referred to healthcare professionals as males. Overall, GPT-4 and GPT-3.5 outperformed all other models. These findings highlight the potential of artificial intelligence in improving access to high-quality maternal health information. Given the complex process of generating high-quality non-English databases, it is desirable to incorporate more accurate translation tools and resourceful architectures for contextualization and customisation.

Keywords Maternal health education, Large language models, Evaluation, Low- and Middle-Income countries

Access to safe and reliable information, as well as the capacity to use it effectively, is critical during pregnancy, birth, and puerperium[1]. With the ultimate goal of improving outcomes, antenatal education is a priority to enhance maternal health literacy and increase women's engagement during pregnancy. In turn, interventions that take advantage of society's exponentially increasing digital resources represent a valuable opportunity to support sustainable development goals for maternal and child health, particularly in developing countries[2].

Despite the challenge of universal accessibility and persisting gender gap, internet use in Low-Middle Income Countries (LMIC) has grown over the last four years[3]. Nevertheless, common online sources of health information often lack content accuracy and suffer from poor accessibility and readability. This can lead to misperceptions about diagnosis, management, and prognosis which may misguide or discourage patients from seeking appropriate care.

[1]Department of Surgery, The Ohio State University Wexner Medical Center and James Comprehensive Cancer Center, Columbus, OH, USA. [2]Department of Gynaecology and Obstetrics, The Ohio State University Wexner Medical Center and James Comprehensive Cancer Center, Columbus, OH, USA. [3]Federal University of Minas Gerais Faculty of Medicine, Belo Horizonte, Brazil. [4]Federal University of Minas Gerais Arts Faculty, Belo Horizonte, Brazil. [5]Federal University of São João Del-Rei Computer Science Department, São João Del-Rei, Brazil. [6]Asenion, Belo Horizonte, Brazil. [7]Henry Ford Hospital, Detroit, MI, USA. [8]Department of Gynaecology and Obstetrics, The Aga Khan University, Karachi, Pakistan. ✉email: vivianresende.ufmg@gmail.com

In this regard, chat-based artificial intelligence (AI) platforms based on Large Language Models (LLM) hold the potential to fundamentally improve healthcare information-seeking mechanisms. LLMs are advanced AI systems trained to understand and generate human-like text, with their applications in healthcare gaining momentum since their introduction in the early 2020s. The safe and responsible deployment of LLMs may provide accurate, reliable, and culturally relevant maternal healthcare information, a critical issue in LMICs[2]. In fact, AI may have the greatest potential benefit in the often resource-limited healthcare systems of LMIC.

Improving maternal health is one of the World Health Organization's key priorities and uneven access to quality maternal health care poses a significant barrier to healthcare equity. In this regard, evaluation and continuous quality improvement are paramount to successfully integrating LLMs in healthcare, ensuring their safety and reliability. To date, despite great enthusiasm, formal assessments of the applicability of LLMs in LMIC healthcare systems and maternal health has not been well-defined. LLMs tailored towards LMIC may function as a source of antenatal education, and digital literacy may enable informed decisions, which are critical for women's empowerment during pregnancy, birth, and puerperium. Although the majority of LLMs are trained on English databases, English is often not the primary language in LMIC. Furthermore, the real-world utility of this technology is influenced by cultural nuances. Therefore, translation tools may play a pivotal role in overcoming this barrier fairly and transparently. Without proper evaluation, LLMs risk spreading harmful or biased information, making assessment crucial for safety and reliability. Identifying key features of a high-performing LLM for LMIC healthcare scenarios could help guide the development and use of this technology in vulnerable communities.

Therefore, the objective of the current study was to assess the potential applicability of LLMs in LMIC healthcare systems. Specifically, we aimed to assess the quality of outputs generated by LLMs pertaining to maternal health. Using a mixed-methods, cross-sectional survey approach, an international panel of obstetrics and gynecologic specialists from Brazil, United States, and Pakistan assessed LLM generated responses in their native languages to a set of questions relating to maternal health. The LLMs' responses were evaluated using metrics for information quality, clarity, readability, and adequacy for the target audience in technical and non-technical domains. As far as we know, this is the first study to evaluate the potential applicability of LLMs as maternal healthcare resources across various languages.

Methods

This cross-sectional survey study adopted a mixed-methods approach, utilizing both quantitative and qualitative evaluation techniques to assess the performance of several LLMs in responding to a series of questions pertaining to maternal health. This study was approved by the Institutional Review Board of the Universidade Federal de Minas Gerais (UFMG), The Aga Khan University and The Ohio State University. We confirm that all research was performed in accordance with relevant guidelines/regulations and Declaration of Helsinki. Informed consent was obtained from all participants. All interactions with the LLM were conducted in compliance with OpenAI's use case policy and The Bill and Melinda Gates Foundation policies[4].

Large language model selection

We compared the performance of four LLMs: GPT-3.5 (OpenAI, Inc. San Francisco, CA),[5] a custom version of GPT-3.5, GPT-4,[6] and Meditron-70B[7] ChatGPT is an LLM based on the GPT architecture developed by OpenAI, and built upon either GPT-3.5 or GPT-4. While the former is freely available to all the users, the latter is an advanced version provided to paid subscribers[8]. Meditron-70B is an open-source medical LLM adapted from Llama 2 to the medical domain. The choice for utilizing GPT-3.5 and GPT-4 relied on their popularity, applicability in a general context, and their training based on vast parameters. Meditron-70B was selected as it is one of the largest LLMs specialized in the medical field, available at the time of this study. Taking advantage of its availability, we also utilized Meditron's training dataset to fine-tune the custom version of GPT-3.5. Custom GPT-3.5 was fine-tuned via OpenAI's proprietary interface using supervised learning with cross-entropy loss, batched inputs, mixed precision, and AdamW optimization. The training corpus included around 48.1B tokens from clinical and biomedical sources, preprocessed as described in the original Meditron pipeline. Hyperparameters (e.g., batch size, learning rate, epochs) and hardware specifications were not disclosed or user-configurable. No reinforcement learning (e.g., RLHF) or parameter-efficient methods (e.g., LoRA) were applied.

Questions and composition

A set of three questions was obtained from a de-identified curated maternal health Q&A database developed by specialists from the UFMG,[9] based on the specialists' clinical experience and relevant topics observed in clinical practice. These were subsequently translated into English and Urdu. All questions were simple and direct based on users' common inputs, aiming to depicting one of each key phases of puerperium (prenatal, labor and nursing) and common topics seen in clinical practice. The questions were:

1. "I already had a C section, can I have a natural birth in my next pregnancy?"
2. "What are my pain relief options during labor and childbirth?"
3. "How many times a day should I breastfeed my baby?"

Once per each question and LLM, questions were submitted to the web chat interfaces of the four above-mentioned LLMs on March 12th 2024 and responses were collected. All questions and generated answers are summarized on Supplementary Table 1.

LLM response generation protocol

A standardized prompt template was developed to guide the response generation of various LLM architectures in a neutral and non-leading way, thus minimizing output bias. This was paramount in maintaining content integrity across the various scenarios presented. The template provided explicit instructions on operating under the persona of a general medical practitioner tasked with addressing medical inquiries in the most comprehensive and informative manner. Specifically, the models were directed to structure their responses using bullet points and paragraphs to enhance readability and clarity. The prompt further stipulated that each response should strive for completeness and be devoid of medical advice, thus focusing solely on providing informational content. Each LLM received identical instructions to ensure uniformity. First-generated responses were always selected to be used in analysis to prevent variability through multiple response regeneration or selection.

The LLMs were engaged in two distinct rounds of response generation. In the first round ("Survey 1"), prompts were input in English (EN-US), and responses were outputted accordingly. These responses were then translated into Portuguese (PT-BR) and Urdu using the Google Translate API to standardize comparative analysis across the target demographics of Brazil, United States, and Pakistan. In that way, evaluators assessed outputs in their native language[10]. The second round ("Survey 2") involved direct prompting in Portuguese, with responses being analyzed exclusively for the Brazilian branch of the study. This was made to evaluate whether responses generated directly in a language other than English differ in quality, completeness, adequacy and clarity compared to those produced in English and later translated. This allowed more granular data on the impact of prompting in LMIC languages. The structure of the LLM Response Generation Protocol is depicted in Supplementary Figs. 1 and 2.

Evaluation

The evaluation was conducted by an international panel comprising 47 obstetrics and gynecologic specialists from Brazil, the United States, and Pakistan. These settings provide geographic and socioeconomic diversity and represent distinct healthcare systems and cultural contexts. Each specialist was presented with a standardized survey that included the three sets of questions & answers for every LLM, resulting in a total of twelve responses per evaluator. Responses were assessed according to information quality, i.e., if the content was correct, complete, and relevant, and clear, i.e., if the response could be easily understood. If an evaluator gave an insufficient score to any of the questions, two more queries were initiated to inquire about the reasons for that score and thus collect more granular data on the rationale behind the assessment. Each metric was evaluated using a five-point Likert scale, ranging from strong disagreement (1) that the criteria were fully met to strong agreement (5). Evaluators were also invited to provide feedback outlining the strengths or weaknesses observed in each LLM response. After concluding their assessment, evaluators were also asked to rank the four LLM answers according to their preference for each question. Finally, a brief section on the applicability of this technology in daily practice also allowed qualitative comments. This structured approach allowed for a comprehensive comparison of the models' abilities to handle a variety of medical queries across different cultural and linguistic contexts.

In order to evaluate the quality, clarity and adequacy of model responses when framed for professional versus general audiences, specialists conducted their evaluation in either a *technical* or *non-technical* domain. Each specialist analysed the adequacy of responses to the target audience they were randomly requested to, judging them as if they were appropriate to be read by domain specialists (audience with technical knowledge) or lay individuals/regular patients (audience without technical knowledge). In aggregate, aiming to capture all the nuances of the two different domains, identify the influence of EN-US vs. PT-BR prompting, as well as evaluate metrics for information quality and clarity in all three languages, a total of eight surveys were created (Supplementary Figs. 1 and 2). To ensure the objectivity and reliability of the evaluation process, each evaluator independently assessed the LLM responses through a standardized online survey designed to elicit detailed scrutiny of the responses' relevance, accuracy, adequacy and comprehensiveness. To safeguard against potential bias from peer influence, evaluators were blinded to the assessments made by their colleagues. Each evaluator was randomly assigned to only respond to one survey, either Survey 1 or Survey 2, and either in the technical or non-technical domain in their native language. This methodological rigor was intended to enhance the validity of the study's findings by reducing subjective variability. All surveys are appended as supplementary material in Annex 1.

Statistical analysis

The Shapiro-Wilk test was conducted and revealed a nonnormal distribution of the data. Therefore, the non-parametric Kruskal-Wallis test was utilized to compare scores between different LLM answers. Continuous variables were presented as medians with inter-quartile range (IQR) and compared using Kruskal-Wallis test. Categorical variables were presented as numbers and percentages and compared with the chi-square test, or Fisher exact test, to evaluate the hypothesis of independence. Descriptive statistics, including means and standard deviation (SD) were computed for all answers in readability analysis. Given the skewed, ordinal nature of the data, intraclass correlation coefficient (ICC) was employed to evaluate inter-rater reliability. Specifically, we used an average rating, fixed-effects consistency ICC model, focusing on the correlation of ratings among evaluators rather than their absolute agreement[11,12]. All tests were 2-sided, $p < 0.05$ was considered statistically significant, and post-hoc adjustments using the Bonferroni correction were applied to adjust for multiple comparisons[13]. Qualitative analysis was performed by reviewing the written feedback provided by the evaluators to identify areas of improvement, such as incorrect, incomplete or inadequate content, as well as notable strengths in LLM´s responses. All statistical analyses were performed using SPSS software version 28.0 (IBM Corporation, Armonk, NY).

Readability analysis

Several metrics such as number of characters, words and sentences were calculated for each LLM-generated response. Furthermore, well stablished readability scores such as the Flesch Reading Ease score and the Flesch-Kincaid Reading Grade level were also calculated. The first provides a decreasing score ranging from 0 to 100, that indicate how difficult a passage in English is to understand, and is a well-established measure of readability[14]. The second is another well recognised metric with a lower grade level suggesting easier readability, and approximates the educational level needed to comprehend a text[15]. Overall, a total of five different readability metrics were included.

Results

Demographic characteristics

Among a total of 47 specialists who fully completed the survey, median age was 50 years (IQR 45–57) and 40 (85%) were female. Overall, the mean clinical experience was 19 years in gynecology and obstetrics and the estimated mean volume of assisted pregnancies per month was 110. A total of 37 (78.7%) specialists from the US (15, 40.5%), Brazil (12, 32.4%) and Pakistan (10, 27.0%) responded to Survey 1, while 10 (21.3%) responded to Survey 2. Overall, respondents of Survey 1 were older and more experienced. Notably, a smaller fraction of specialists from the US had postgraduate academic degrees (postgraduate academic degrees; US, 14.3% vs. BR, 75.0% vs. PK, 70.0%; $p < 0.001$); other variables were evenly distributed among specialists' groups. Table 1 details the demographics of the two survey populations stratified by language.

Quantitative

Overall, in the non-technical assessment domain, the mean score was 4.03/5 for GPT-3.5, 2.63/5 for custom GPT-3.5, 3.87/5 for GPT-4, and 3.35/5 for Meditron-70b. Comparatively, in the technical assessment scenario, the overall score was 3.82/5 for GPT-3.5, 2.77/5 for custom GPT-3.5, 4.05/5 for GPT-4 and 3.62/5 for Meditron-70b. In aggregate, generated responses garnered relatively high scores for clarity (Q&A-1 3.5, Q&A-2 3.7, Q&A-3 3.8) and for quality of content (Q&A-1 3.2, Q&A-2 3.2, Q&A-3 3.7). Notably, the most common limitation to quality was incomplete content. Interestingly, overall scores varied relative to the questions; for Q&A-1 and − 2, GPT-3.5 and GPT-4 demonstrated the highest overall scores, while for Q&A-3, Meditron-70b had the highest overall scores, particularly in Portuguese and Urdu. Based on average ratings, two-way fixed-effects consistency model demonstrated an overall ICC of 0.94 (95% CI: 0.88–0.97) for Survey 1 (average rating k = 37) and 0.78 (95% CI: 0.62–0.89) for Survey 2 (average rating k = 10), indicating excellent inter-rater reliability. Similarly, Survey 1 analysis by language demonstrated an ICC of 0.92 (95% CI: 0.85–0.97) for English (average rating k = 15), 0.86 (95% CI: 0.77–0.93) for Portuguese (average rating k = 12) and 0.79 (95% CI: 0.64–0.90) for Urdu (average rating k = 10).

Survey 1 scores attributed to answers generated by GPT-3.5 and GPT-4 were consistently higher (Overall, GPT-3.5, 3.9 (3.8–4.1); GPT-4.0, 3.9 (3.8–4.1); Custom GPT-3.5, 2.7 (2.5–2.8); Meditron-70b, 3.5 (3.3–3.6), $p = 0.000$*; English, GPT-3.5, 4.1 (3.9–4.3); GPT-4.0, 4.2 (4.0-4.3); Custom GPT-3.5, 2.9 (2.6–3.1); Meditron-70b, 3.6 (3.4–3.8), $p = 0.000$*; Portuguese, GPT-3.5, 4.2 (4.0-4.4); GPT-4.0, 3.9 (3.6–4.2); Custom GPT-3.5, 2.8

Variables	Survey 1				Survey 2
	English *n* = 15 (40.5%)	Portuguese *n* = 12 (32.4%)	Urdu *n* = 10 (27.0%)	p-value	Portuguese *n* = 10 (100%)
Age, years				0.606	
Median (IQR)	51 (45–57)	48 (41–55)	51 (45–57)		37 (29–46)
Gender				0.332	
Male	3 (20.0%)	2 (16.7%)	0 (0.0%)		1 (10.0%)
Female	12 (80.0%)	10 (83.3%)	10 (100%)		9 (90.0%)
Specialty				0.250	
OB/GYN	15 (100%)	12 (100%)	9 (90.0%)		9 (90.0%)
Midwife	0 (0.0%)	0 (0.0%)	1 (10.0%)		1 (10.0%)
Time in specialty (year)				0.857	
Median (IQR)	20 (14–26)	18 (11–26)	19 (12–25)		8.9 (1–19)
Pregnant assisted monthly				0.053	
Median (IQR)	90 (46–134)	72 (15–129)	178 (81–275)		40 (19–58)
Postgraduate academic degrees				< 0.001*	
No	12 (85.7%)	3 (25.0%)	3 (30.0%)		6 (60.0%)
Yes	2 (14.3%)	9 (75.0%)	7 (70.0%)		4 (40.0%)
Theme of survey				0.924	
Technical	7 (46.7%)	5 (41.7%)	5 (50.0%)		5 (50.0%)
Non-technical	8 (53.3%)	7 (58.3%)	5 (50.0%)		5 (50.0%)

Table 1. Demographic characteristics of evaluators. IQR, interquartile range; OB/GYN, Obstetrician/Gynecologist; * signify statistical significance ($p < 0.05$).

(2.5–3.1); Meditron-70b, 3.6 (3.3–3.9), $p = 0.000$*; Urdu, GPT-3.5, 3.4 (3.1–3.6); GPT-4.0, 3.5 (3.3–3.8); Custom GPT-3.5, 2.3 (2.0-2.5); Meditron-70b, 3.1 (2.8–3.4), $p = 0.000$*) (Fig. 1). A comparison of quality and clarity scores between technical and non-technical domains demonstrated no significant differences. Interestingly, the performance of Meditron-70b was similar to that of GPT-3.5 and GPT-4 for Q&A-3 (Table 2). In contrast, comparing quality and clarity between languages, there were significant differences. For example, regarding the quality of answers, some LLMs had better performance in English [Q&A-2 Meditron-70b, 4.0 (3.0–4.0) EN-US vs. 2.5 (2.0–3.0) PT-BR vs. 2.5 (2.0–3.0) Urdu, $p = 0.015$*; Q&A-3 GPT-3.5, 4.0 (4.0–4.0) EN-US vs. 4.0 (4.0–5.0) PT-BR vs. 3.0 (2.0–4.0) Urdu, $p = 0.004$*]. Notably this pattern changed when analysing the clarity of answers, and clarity increased for Portuguese, although the asymmetry between English and Urdu remained [Q&A-3; GPT-3.5, 4.0 (4.0–5.0) EN-US vs. 5.0 (4.0–5.0) PT-BR vs. 2.5 (2.0–4.0) Urdu, $p < 0.001$*; GPT-4.0, 4.0 (4.0–5.0) EN-US vs. 5.0 (4.5-5.0) PT-BR vs. 4.0 (3.0–4.0) Urdu, $p = 0.006$*] (Table 3). Supplementary Table 2 provides more granular data on the comparison of quality and clarity between languages included in Survey 1 by domains. Quantitative evaluation of suboptimal content in Survey 1 by languages demonstrated high percentages of incomplete content across all languages and relatively higher percentage of incorrect content perceived among Portuguese evaluators (Supplementary Table 3).

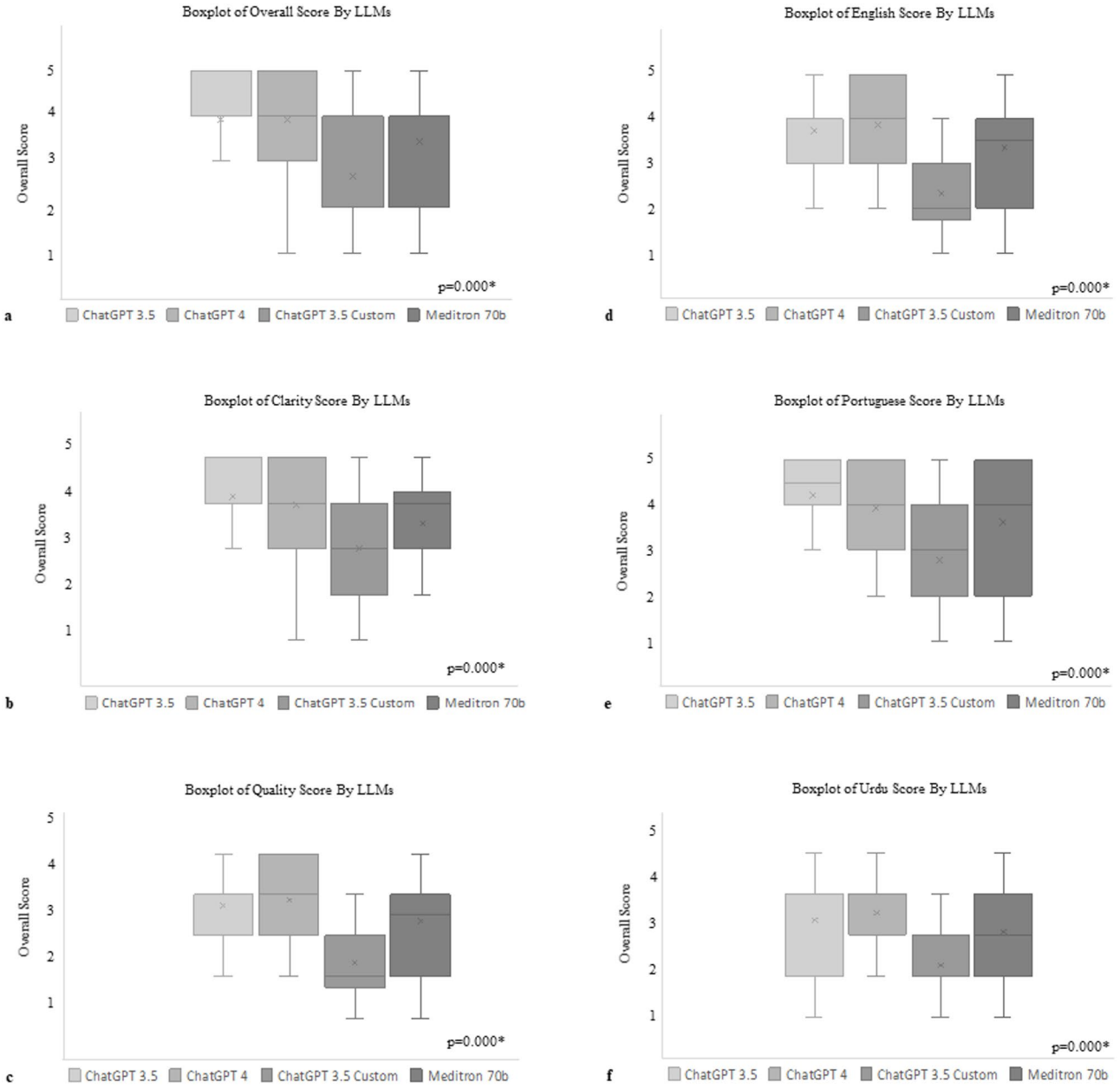

Fig. 1. Boxplots depicting average overall evaluator responses of Survey 1 by models, including: (a) Overall; (b) Clarity; (c) Quality; (d) English; (e) Portuguese and (f) Urdu. Scores range from 1 to 5, with higher scores indicating better overall performance. LLM, Large Language Model, * signify statistical significance ($p < 0.05$).

Vignettes	Quality (IQR)		*p*-value	Clarity (IQR)		*p*-value
	Non-technical	Technical		Non-technical	Technical	
Question 1						
ChatGPT 3.5	4.0 (2.5–4.5)	4.0 (4.0-4.5)	0.232	4.0 (4.0–5.0)	4.5 (4.0–5.0)	0.141
ChatGPT 4.0	4.0 (3.0–5.0)	4.0 (2.5-4.0)	0.326	4.0 (3.0–5.0)	4.0 (2.5-4.0)	0.442
ChatGPT 3.5 Custom	2.0 (1.0–3.0)	2.0 (2.0–3.0)	0.537	2.0 (2.0-3.5)	3.5 (2.0–4.0)	0.110
Meditron 70b	2.0 (2.0–3.0)	3.0 (2.0–4.0)	0.970	3.0 (2.0–4.0)	4.0 (2.5-4.0)	0.238
Question 2						
ChatGPT 3.5	4.0 (2.0–5.0)	4.0 (2.0–4.0)	0.598	4.0 (4.0–5.0)	4.5 (4.0–5.0)	0.497
ChatGPT 4.0	4.0 (4.0–5.0)	4.0 (3.0-4.5)	0.297	4.0 (3.5-5.0)	4.0 (3.5-5.0)	0.775
ChatGPT 3.5 Custom	2.0 (1.0–3.0)	2.0 (1.0–3.0)	0.707	3.0 (2.0–4.0)	3.5 (2.0–4.0)	0.964
Meditron 70b	3.0 (2.0–4.0)	3.0 (2.5-4.0)	0.232	3.5 (2.0–4.0)	3.5 (2.0–4.0)	0.497
Question 3						
ChatGPT 3.5	4.0 (3.5–4.5)	4.0 (4.0–4.0)	0.987	4.0 (3.5-5.0)	4.0 (3.5-5.0)	0.789
ChatGPT 4.0	4.0 (3.5-5.0)	4.0 (4.0-4.5)	0.765	4.5 (4.0–5.0)	4.0 (4.0–5.0)	0.519
ChatGPT 3.5 Custom	2.0 (2.0-3.5)	2.0 (2.0-3.5)	0.888	3.0 (2.0–4.0)	3.0 (2.0–4.0)	0.814
Meditron 70b	4.0 (4.0–5.0)	4.0 (4.0–5.0)	0.863	4.0 (4.0–5.0)	4.0 (3.0–5.0)	0.741

Table 2. Comparison of scores of quality and clarity between technical and non-technical themes of survey 1 (median). IQR, interquartile range; * signify statistical significance ($p < 0.05$).

Vignettes	Quality (IQR)			*p*-value	Clarity (IQR)			*p*-value
	English	Portuguese	Urdu		English	Portuguese	Urdu	
Question 1								
ChatGPT 3.5	4.0 (4.0–5.0)	4.0 (3.5-5.0)	4.0 (2.0–4.0)	0.349	4.0 (4.0–5.0)	5.0 (4.0–5.0)	4.0 (3.0–4.0)	0.012*
ChatGPT 4.0	4.0 (3.0–5.0)	3.0 (2.0-4.5)	4.0 (2.0–4.0)	0.315	4.0 (4.0–5.0)	2.5 (2.0–5.0)	4.0 (3.0–4.0)	0.150
ChatGPT 3.5 Custom	2.0 (2.0–4.0)	2.0 (1.5-3.0)	2.0 (2.0–2.0)	0.617	3.5 (2.0–4.0)	3.0 (2.0–4.0)	2.0 (1.0–3.0)	0.072
Meditron 70b	2.5 (2.0–4.0)	3.0 (2.0-4.5)	2.0 (2.0–3.0)	0.303	3.0 (2.0–4.0)	4.0 (3.0–5.0)	2.0 (2.0–4.0)	0.072
Question 2								
ChatGPT 3.5	3.0 (2.0–4.0)	4.0 (3.0–5.0)	4.0 (4.0–4.0)	0.324	4.0 (4.0–5.0)	5.0 (4.5-5.0)	4.0 (3.0–4.0)	0.054
ChatGPT 4.0	4.0 (4.0–5.0)	4.0 (2.0–5.0)	4.0 (3.0–5.0)	0.506	4.0 (4.0–5.0)	4.5 (3.5-5.0)	4.0 (2.0–5.0)	0.353
ChatGPT 3.5 Custom	2.0 (1.0–3.0)	2.0 (1.0–3.0)	2.5 (1.0–3.0)	0.783	4.0 (3.0–4.0)	3.5 (3.0–4.0)	2.0 (2.0–3.0)	0.059
Meditron 70b	4.0 (3.0–4.0)	2.5 (2.0–3.0)	2.5 (2.0–3.0)	0.015*	4.0 (3.0–4.0)	3.5 (2.0-4.5)	2.0 (2.0–4.0)	0.073
Question 3								
ChatGPT 3.5	4.0 (4.0–4.0)	4.0 (4.0–5.0)	3.0 (2.0–4.0)	0.004*	4.0 (4.0–5.0)	5.0 (4.0–5.0)	2.5 (2.0–4.0)	< 0.001*
ChatGPT 4.0	4.0 (4.0–5.0)	5.0 (4.0–5.0)	4.0 (2.0–4.0)	0.077	4.0 (4.0–5.0)	5.0 (4.5-5.0)	4.0 (3.0–4.0)	0.006*
ChatGPT 3.5 Custom	2.0 (2.0–4.0)	2.0 (2.0-3.5)	2.0 (2.0–3.0)	0.843	3.0 (2.0–4.0)	3.0 (3.0-4.5)	2.0 (2.0–3.0)	0.029*
Meditron 70b	4.0 (4.0–4.0)	5.0 (4.0–5.0)	4.0 (4.0–5.0)	0.216	4.0 (3.0–5.0)	5.0 (4.0–5.0)	4.0 (3.0–5.0)	0.162

Table 3. Comparison of scores of quality and clarity between languages of survey 1 (median). IQR, interquartile range; * signify statistical significance ($p < 0.05$).

Survey 2 also had higher GPT-3.5 and GPT-4 overall scores (Overall, GPT-3.5, 4.7 (4.6–4.9); GPT-4.0, 4.4 (4.2–4.7); Custom GPT-3.5, 3.4 (3.0-3.8); Meditron-70b, 4.1 (3.8–4.4), $p = 0.000$*) with clarity being high across all Q&As and LLMs. Regarding the quality of answers, GPT-3.5 and GPT-4 consistently outperformed custom GPT-3.5 and Meditron-70b. Interestingly, a comparison of quality and clarity between technical and non-technical themes demonstrated no significant differences (Supplementary Table 4). Similar to Survey 1, quantitative valuation of suboptimal content of Survey 2 demonstrated high percentages of incomplete content, followed by relatively high percentages of incorrect content among evaluators (Supplementary Table 5).

Qualitative

Qualitative analyses varied greatly between LLMs, languages and questions, and focused on reasons for not attributing the highest score of agreement relative to quality and clarity. When analyzed by specialists from the US, GPT-3.5 and GPT-4 were able to effectively produce good answers to all questions in a fairly clear manner. However, answers often lacked detailed information. Specifically, evaluators noted that many patients would like to know more and that it would be helpful to provide additional resources for information to be discussed with their providers. Furthermore, evaluators pointed out some outdated terms used by LLMs no longer seen on current practice, such as "VBAC" (vaginal birth after cesarean), currently substituted by the more appropriate "TOLAC" (trial of labor after cesarean). Moreover, some answers lacked enough description to provide nuanced or site-specific discussion. For example, strategies for pain relief during labor may or may not be available depending on the hospital resources and policy. In contrast, Meditron-70b was able to effectively answer the questions, with good information and easily understood language, especially regarding Question-3. It was noted, however, that there was room for improvement in the content. For example, the lack of inclusion of medical therapies in choices of pain relief during childbirth, which highlights the need for having a human-in-the-loop considering the models' current limitations.

In Portuguese, specialists pointed out that GPT-3.5 sometimes failed to provide complete information and even omitted potential risk. For example, not clarifying the possibility of "uterine rupture" in vaginal birth after cesarean or not including epidural analgesia as a strategy for pain relief during childbirth. For GPT 4.0, besides mentioning the need for more complete answers, such as "listing favorable and unfavorable factors for VBAC", translation errors were often mentioned. Poor sentence construction and mistranslation diminished the authenticity and clarity of information. Again, despite being clear, answers generated by custom GPT-3.5 were considered "generic and superficial", and specialists noted the need for a more complete answer. Meditron-70b responses were considered generic and incomplete; for example, by only addressing the risks and not providing any information about benefits of certain procedures. Evaluators criticized the model's apparent negative view of labor/delivery, assuming that "there will always be pain that will require some method of relief", and its failure to address the fact that each woman may react differently according to their tolerance. Theses biases regarding the inevitability of painful labour and exclusion of medical therapies may be reflecting an existing priority within the healthcare sector that is being reflected back by the LLMs.

In Urdu, answers generated by GPT-3.5 and GPT-4 were generally considered adequate, but sometimes extra information was desired. Custom GPT-3.5 responses lacked sufficient information, overlooked details and omitted important points. In contrast, answers generated by Meditron-70b were often considered very good, containing pertinent and detailed information. However, evaluators noted that Meditron-70b struggled relative to information about the risk of mortality in certain scenarios. Evaluators noted that crude information about the risk of death was generated in a "scary way", without mentioning the available means to mitigate it, and may mislead patients eligible for vaginal birth. Importantly, throughout all LLMs, most criticism regarding responses was related to the apparent poor Urdu translation and the urge to improve it. All results from qualitative analyses are summarized on Supplementary Table 6.

Readability

Readability analysis in Portuguese demonstrated an even distribution of scores and characteristics across the 3 Q&As. The mean character count for the responses was 910.5 (SD, 738.2 Q&A-1), 840.3 (SD, 760.8 Q&A-2), and 828.5 (SD, 659.2 Q&A-3), with a mean sentence count of 10.0 (SD, 8.8 Q&A-1), 11.5 (SD, 11.1 Q&A-2) and 8.8 (SD, 6.2 Q&A-3). The average Flesch Reading Ease score for the LLM responses was 48.8 (SD, 6.8 Q&A-1), 46.6 (SD, 7.6 Q&A-2) and 51.6 (SD, 8.0 Q&A-3), representing a difficult to fairly difficult reading level. The average Flesch-Kincaid Reading Grade level was 11.9 (SD, 0.8 Q&A-1), 11.5 (SD, 1.8 Q&A-2) and 11.0 (SD, 1.3 Q&A-3), indicating language typical of college-level texts. Similarly, readability analysis in English had a mean character count of 514, (SD, 323.9 Q&A-1), 569.5 (SD, 536.9 Q&A-2), and 534.0 (SD, 389.3 Q&A-3), with a mean sentence count of 5.3 (SD, 4.3 Q&A-1), 11.8 (SD, 6.8 Q&A-2) and 6.3 (SD, 5.5 Q&A-3). The average Flesch Reading Ease score was 40.7 (SD, 24.5 Q&A-1), 50.2 (SD, 7.9 Q&A-2) and 55.6 (SD, 10.3 Q&A-3) and the average Flesch-Kincaid Reading Grade level was 13.1 (SD, 5.1 Q&A-1), 9.4 (SD, 1.5 Q&A-2) and 9.9 (SD, 2.0 Q&A-3). All these values indicate that the responses may require a fairly high reading level for comprehension. Moreover, gender bias was detected in that models referred to healthcare professionals as males. All values are summarized in Supplementary Table 7.

Discussion

Despite significant progress over the past two decades, full access to quality maternal healthcare and equitable perinatal outcomes remain challenging. This challenge is particularly apparent in LMIC. In fact, among the 287,000 maternal deaths during or after pregnancy globally, approximately 95% occurred in LMIC with the majority being preventable[16]. In this regard, equitable access to reliable health-related information is critical during the prenatal, labor and the postnatal periods[1]. Digitization efforts over the previous two decades have led to the rapid proliferation of healthcare information resources. Furthermore, with the advent of publicly available

chat-based AI platforms, information on a wide variety of topics can be obtained in a conversational setting. Theoretically, chat-based AI platforms may be a critical tool in the effort to democratize access to healthcare information, particularly in LMIC that face significant resource limitations in terms of healthcare centres as well as personnel. The enthusiasm for LLM is tempered, however, by the dangers of potentially inaccurate and unsafe healthcare information. In this regard, the current study sought to assess the performance of several LLMs as maternal health information resources using an international panel of experts. Furthermore, by assessing their performance in English, Portuguese and Urdu, this study is the first to evaluate the potential applicability of LLMs as maternal healthcare resources across various languages. Although responses to common maternal health questions were characterized by generally high scores for clarity and quality of content, readability and poor translation were identified as key areas of improvement. The rapid growth of LLMs has outpaced its scientific assessment and this study represents a step towards the identification of key features of a high performing LLM for LMIC healthcare scenarios, which could help guide the development and use of this technology in vulnerable communities.

Publicly available chat-based AI platforms have captured the imagination of the public since the launch of ChatGPT in November 2022. Their applicability as a source of medical information for patients in various fields has been a key area of research[17–20]. The present study demonstrated high scores for clarity (Q&A-1 3.5, Q&A-2 3.7, Q&A-3 3.8) and quality of content (Q&A-1 3.2, Q&A-2 3.2, Q&A-3 3.7) in maternal health. Notably, there were no differences in scores in the technical and nontechnical domains. A possible explanation for this may be related to the readability analysis as a majority of the responses required a high school to college level of comprehension. This fairly high educational level for comprehension can be attributed to the type of text with which models are trained and the absence of specific prompt design. Furthermore, translation might have impacted Portuguese readability scores, as certain language choices are less common in Portuguese than in English, as is the case of passive voice and other grammatical structures. Moreover, LLMs tend to produce language closer to written texts in which they were trained, and English readability scores might reflect the standard data utilized by the LLMs. Notably, the American Medical Association recommends patient-facing information to be at a sixth grade or lower level of reading[21]. Recent studies have demonstrated that incorporating the required level of reading (sixth grade or lower) in the LLM prompt may be critical to produce responses at an appropriate reading level[21]. This was not done in the current study; for appropriate use, the public may have to be educated on the appropriate way to prompt these technologies. In the future, LLMs custom-built to provide health information to patients should incorporate algorithms to ensure that responses are presented at a sixth grade reading level or lower. Furthermore, LLM responses may incorporate customized illustrations and videos to better explain concepts in maternal health; the use of visual representations has been demonstrated to be associated with improved comprehension[22].

To our knowledge, this study represents the first analysis of the performance of various LLMs on an international scale relative to pregnant maternal health literacy. Notably, the vast majority of LLMs are trained and presented in English. As the premise of LLM usage in the healthcare sector is to increase accessibility to medical information, it is critical to assess their performance in various native languages. This point is emphasized by the fact that 65% of all internet users prefer to receive information on the internet in their native languages[23]. In the current study, we incorporated Brazil and Pakistan as study sites and used Portuguese and Urdu as the languages of interest, alongside English. An interesting finding was that there were largely no differences in the quality of responses among the three languages, with the only exception being Meditron for Q&A 2 and ChatGPT-3.5 for Q&A 3. As these LLMs were initially trained in English, this finding may signify that the additional step of translation did not appear to affect the quality of content. However, this difference did significantly impact the clarity of responses in both the technical and nontechnical domains, especially for Urdu. This finding was particularly evident in the qualitative analysis, with poor sentence construction and mistranslation frequently cited as significant barriers for both Portuguese and Urdu. These data highlight the current limitations in applying these technologies on a global scale. A potential solution may lie in the development of more sophisticated built-in neural-based translation tools, which do not simply translate prompts but contextualize them in a conversational manner in a larger number of languages. Furthermore, evaluators noticed the presence of gender bias, in that the models always referred to healthcare professionals as males. This risks reaffirming the medical sector as being male-dominated, harming the platforms accessibility for women and non-binary medical professionals, and potentially having a larger impact in LMICs where midwives, the front line healthcare providers for pregnant women, are predominantly women. There are many possible sources of biases, such as data-related, human-induced and machine generated biases, and diverse sources of biases can perpetuate one another. Understanding how diverse biases affect AI systems and recommendations is critical. Given that different sources of biases add to one another exponentially, researchers and medical personnel should employ possible safeguards with a "bias-in-mind" approach[24]. Finally, the application of these tools at the training stage may lead to better results, rather than simply at the back end for response generation.

Our findings may help design future LLMs custom-built to cater to the needs of patients who seek to gain an understanding of maternal health-related issues and possibly assist front-line workers. LLMs need to incorporate a variety of different subspecialities, with the ultimate aim to develop an all-encompassing medical LLM. In the present study, ChatGPT 3.5 and 4.0 significantly outperformed the custom ChatGPT 3.5 as well as Meditron-70B. Of note, ChatGPT 3.5 is trained on 175 billion parameters while the 4.0 version is possibly trained on 1.7 trillion parameters[5,6]. In comparison, Meditron-70b is trained on 70 billion,[7] while the custom ChatGPT 3.5 was also fine-tuned on the same set of medical information as Meditron 70b. Furthermore, ChatGPT has undergone several iterations with exponentially increasing sophistication. The impact of fine tuning large closed-source LLMs, like ChatGPT 3.5, demonstrated poor results, even when utilizing large databases with positive previous results on open-sourced LLMs, such as Meditron 70B - based on Llama 2. This may relate to the quality and format of information available for training and fine-tuning, even though we utilized OpenAI's Fine-Tuning

tool[25]. The capacity of ChatGPT to articulate texts and communicate had a greater influence on the perceived quality over domain specific trained LLMs, which highlight an interesting pathway for LLM development and improvement. Given that our findings suggested that non-trained LLMs outperformed adapted models, we believe that more complex architectures such as Retrieval Augmented Generation (RAG) may be a promising avenue to design domain-specific LLMs that are contextually relevant to LMICs. That is because this architecture framework expands on the concept of vector search and combines LLMs with information retrieval systems to generate more accurate and relevant text. Moreover, attention should also be directed towards the quality of the translation tools available, given the complex and time-consuming process of generating high-quality training datasets in non-English languages. In all, the exciting potentials that LLMs offer to make maternal healthcare more accessible for all can only be realised if they are customised to address context-specific needs. Therefore, instead of diverting our efforts towards building large training datasets from scratch for non-English languages, we should focus on optimising translation, contextualization and customisation strategies that build upon existing datasets. Nevertheless, it is also important to understand that while LLMs can generate coherent and structured responses, they should not be considered stand-alone decision-making tools in healthcare. Their limitations in clinical reasoning and real-world applicability further highlight the need for scientific research to assess its capabilities and continuous human supervision and expert validation.

This study has several limitations. Although the selection of questions did encompass all key phases of puerperium, they are limited in representing all possible interactions between users and the LLMs regarding maternal healthcare. Moreover, the Q&A approach to the LLMs may not reflect real-word interaction scenarios. Despite the large number and diversity of evaluators, their distribution was uneven across countries and it may not be representative of all maternal healthcare specialists, for all different languages and settings. Due to probabilistic text generation, responses may vary even with identical prompts. While our study standardized evaluation by selecting the first response, future research should assess variability through multiple iterations and use statistical methods to quantify response consistency and reliability. Furthermore, LLMs are rapidly evolving, which may harm the reproducibility of this study and its long-term relevance. Assessing the impact of LLM outputs directly on all audiences with varying levels of health literacy would also be a valuable extension of this work. Unfortunately, hyperparameters such as batch size, learning rate, and number of epochs, as well as hardware specifications (e.g., GPU model, memory), are not accessible or configurable through the OpenAI API. Finally, future research should go beyond subjective quality assessment and incorporate objective measures, such as factual accuracy validation, bias detection, and external clinical guideline adherence, to better understand the safety and reliability of AI-generated maternal health content.

In conclusion, this study is the first to analyse and compare the performance of various LLMs on an international scale in maternal health education. The data provide a strong foundation to improve quality of LLMs in maternal healthcare settings for non-English languages and their variations. Overall, LLM responses were characterized by high scores for clarity and quality of content, however readability and poor translation were identified as key areas of improvement. Furthermore, clarity of content was notably better in English compared with Urdu and Portuguese, pointing to content being "lost in translation." In all, this study highlights the exciting possibilities for the application of LLMs to improve medical literacy in maternal health. However, the need for significant monetary investment, the incorporation of more accurate translation tools and resourceful architectures for contextualization and customisation, given the poor availability of large non-English medically oriented databases represent key obstacles to overcome. While clarity, readability, quality and adequacy all enhance accessibility to more reliable content, they should not be interpreted as proxies for medical accuracy. Although LLMs hold potential as supplementary educational tools, their application in clinical practice should be approached with caution until further validation studies are conducted, taking advantage of interdisciplinary collaborations to establish robust frameworks for evaluating and monitoring AI-generated health information. This work may shed light on future pathways for tailoring LLMs for resource-poor settings within LMICs – especially on the many considerations around subcultures, linguistics, and socio-demographics that go into building effective LLMs.

Data availability

Data availability statementAs leader of this study I state that: The datasets used and analysed during the current study available from the corresponding author on reasonable request.

Received: 22 January 2025; Accepted: 20 May 2025
Published online: 02 July 2025

References

1. Renkert, S. & Nutbeam, D. O. N. Opportunities to improve maternal health literacy through antenatal education: an exploratory study. *Health Promot. Int.* **16** (4), 381–388 (2001).
2. Jafree, S. R., Bukhari, N., Muzamill, A., Tasneem, F. & Fischer, F. Digital health literacy intervention to support maternal, child and family health in primary healthcare settings of Pakistan during the age of coronavirus: study protocol for a randomised controlled trial. *BMJ Open.* **11** (3), e045163 (2021).
3. Domicílios, T. I. C. Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros (2019). https://cetic.br/pt/pesquisa/domicilios/publicacoes/.
4. OpenAI. Terms & policies. https://openai.com/policies (2024).
5. OpenAI. ChatGPT (version 3.5), 2024. (2023). https://www.openai.com/.
6. OpenAI. ChatGPT (version 4), 2024. (2024). https://www.openai.com/.
7. Hugo, T. et al. Llama 2: Open foundation and fine-tuned chat models (EPFL, 2024).
8. OpenAI. What is ChatGPT? https://help.openai.com/en/articles/6783457-what-is-chatgpt (2024).

9. Moraes Carrilho, J. et al. Pregnant users' perceptions of the birth plan interface in the my prenatal care app: observational validation study. *JMIR Form. Res.* **3** (1), e11374. https://doi.org/10.2196/11374 (2019).
10. Castilho, S., Mallon, C., Meister, R. & Yue, S. *Do Online Machine Translation Systems Care for Context??* (What about a GPT model?, 2023).
11. Hallgren, K. A. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor. Quant. Methods Psychol.* **8**, 23–34 (2012).
12. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods*. **1**, 30–46 (1996).
13. Bland, J. M. & Altman, D. G. Multiple significance tests: the bonferroni method. *BMJ* **310**, 170 (1995).
14. FLESCH R. A new readability yardstick. *J. Appl. Psychol.* **32**, 221–233 (1948).
15. Kincaid, J. P. et al. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel (1975).
16. WHO, World Health Organization (2024, 31 Jul 2024). news-room/fact-sheets/detail/maternal-mortality, https://www.who.int/news-room/fact-sheets/detail/maternal-mortality.
17. Moazzam, Z. et al. A paradigm shift: online artificial intelligence platforms as an informational resource in bariatric surgery. *Obes. Surg.* **33** (8), 2611–2614. https://doi.org/10.1007/s11695-023-06675-3 (2023).
18. Bellamkonda, N. et al. Evaluating the accuracy of ChatGPT in common patient questions regarding HPV + Oropharyngeal carcinoma. *Ann. Otol. Rhinol. Laryngol.* https://doi.org/10.1177/00034894241259137 (2024).
19. Belge Bilgin, G. et al. Performance of ChatGPT-4 and Bard chatbots in responding to common patient questions on prostate cancer 177Lu-PSMA-617 therapy. *Front Oncol.* **14**, 1386718. https://doi.org/10.3389/fonc.2024.1386718 (2024).
20. Musheyev, D. et al. Readability and information quality in Cancer information from a free vs paid chatbot. *JAMA Netw. Open.* **7** (7), e2422275. https://doi.org/10.1001/jamanetworkopen.2024.22275 (2024).
21. Weiss, B. D. Health literacy. *Am. Med. Assoc.* **253**, 358 (2003).
22. Mbanda, N., Dada, S., Bastable, K., Ingalill, G. B. & Ralf, W. S. A scoping review of the use of visual aids in health education materials for persons with low-literacy levels. *Patient Educ. Couns.* **104** (5), 998–1017 (2021).
23. Research, C. S. A. (2024, accessed 8 Jan 2024). https://csa-research.com/Blogs-Events/CSA-in-the-Media/Press-Releases/Consumers-Prefer-their-Own-Language.
24. Perets, O. et al. Inherent bias.in electronic health records: a scoping review of sources of bias. *MedRxiv*. https://doi.org/10.1101/2024.04.09.24305594 (2024).
25. OpenAI's Fine-Tuning tool. https://platform.openai.com/docs/guides/fine-tuning (2025).

Acknowledgements

Bill & Melinda Gates Foundation for financial support by AI Grand Challenges (grant number INV-062615).

Author contributions

Authors contribution: Henrique A. Lima contributed with Study Conception, Data collection, Data analysis, Writing and Manuscript review. Pedro H. F. S. Trocoli-Couto contributed with Study Conception, Data collection, Data analysis. Writing and Manuscript review. Zorays Moazzam contributed with Study Conception, Data collection, Data analysis, Writing and Manuscript review. Leonardo C. D. Rocha contributed with Data collection, AI model´s training and development, Writing and Manuscript review. Adriana Pagano contributed with Data analysis, AI model´s training and development, Writing and Manuscript review. Felipe F. Martins contributed with Data collection, AI model´s training and development, Writing and Manuscript review. Lucas T. Brabo contributed with Data collection, AI model´s training and development, Writing and Manuscript review. Zilma S. N. Reis contributed with Data collection, Writing and Manuscript review. Lisa Keder contributed with Data collection, Writing and Manuscript review. Aliya Begum contributed with Data collection, Writing and Manuscript review. Marcelo H. Mamede contributed with Data analysis, Writing and Manuscript review. Timothy M. Pawlik contributed with Study Conception, Writing and Manuscript review. Vivian Resende contributed with Study Conception, Writing and Manuscript review.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-03501-x.

Correspondence and requests for materials should be addressed to V.R.